NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

ESTIMATING THE DEPTH OF THE NAVY RECRUITING MARKET

by

Emilie M. Monaghan

September 2016

Thesis Advisor:	Lyn R. Whitaker
Second Reader:	Jonathan K. Alt

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2016	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE ESTIMATING THE DEPTH OF THE NAVY RECRUITING MARKET	**5. FUNDING NUMBERS**
6. AUTHOR(S) Emilie M. Monaghan	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) Navy Recruiting Command, N5	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

This research develops a statistical model for predicting the number of leads, as an indicator of market depth, that a ZIP code will produce for Navy Recruiting Command (NRC). The U.S. Navy recruits from all over the country, using previous accessions in a recruiting district to assign recruiters and goals. This research develops statistical models to determine the key drivers of the number of leads at the ZIP code level. This research develops a Poisson regression model to predict the number of leads using factors such as IRS-estimated population size and five cluster membership factors constructed from publicly available data sources. We recommend that NRC make use of the Poisson regression model in order to determine high-yield ZIP codes for market depth.

14. SUBJECT TERMS Navy Recruiting Command, NRC, recruiting, data analysis, clusters, generalized linear regression, leads, Poisson regression	**15. NUMBER OF PAGES** 87
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

ESTIMATING THE DEPTH OF THE NAVY RECRUITING MARKET

Emilie M. Monaghan
Captain, United States Marine Corps
B.S., United States Naval Academy, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2016

Approved by: Lyn R. Whitaker
Thesis Advisor

LTC Jonathan K. Alt
Second Reader

Patricia A. Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This research develops a statistical model for predicting the number of leads, as an indicator of market depth, that a ZIP code will produce for Navy Recruiting Command (NRC). The U.S. Navy recruits from all over the country, using previous accessions in a recruiting district to assign recruiters and goals. This research develops statistical models to determine the key drivers of the number of leads at the ZIP code level. This research develops a Poisson regression model to predict the number of leads using factors such as IRS-estimated population size and five cluster membership factors constructed from publicly available data sources. We recommend that NRC make use of the Poisson regression model in order to determine high-yield ZIP codes for market depth.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ACS	Army Custom Segments
AIC	An Information Criterion
CENSUS	United States Census Bureau
CHSI	Community Health Status Indicators
CONUS	Continental United States
DOD	Department of Defense
FIPS	Federal Information Processing Standard
FY	Fiscal Year
GAO	Government Accountability Office
HUD	Department of Housing and Urban Development
IRS	Internal Revenue Service
LRT	Likelihood Ratio Test
NRC	Navy Recruiting Command
Pam	Partitioning around Medoids
P.O.	Post Office
PRIZM NE	Potential Rating Index for Zone Improvement Plan Code Markets New Evolution
QMA	Qualified Military Available
Q-Q	Quantile- Quantile
RAF	Recruiter Allocation Factor
USAREC	United States Army Recruiting Command
USPS	United States Postal Service
ZCTA	ZIP Code Tabulation Area
ZIP	Zone Improvement Plan

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The United States Navy recruits from all over the country, using previous accessions to assign recruiters and goals. As resources available to the Navy decrease, as well as the Navy's end strength, recruiting must be done more efficiently. This research develops a statistical model for predicting the number of national Navy active duty leads, as an indicator of market depth, that a ZIP code will produce for Navy Recruiting Command (NRC).

In order to remove any possible duplicates between local and national leads, and to mitigate the influence of local recruiting efforts, we remove local leads, which account for approximately 7% of all leads. The training set utilizes national leads data from Fiscal Year (FY) 11 to FY 14. We aggregate leads by year and Zone Improvement Plan (ZIP) code to construct a four-year total number of leads for the training set for each ZIP code. Number of leads from FY 15, which only comprise approximately 11% of FY 11 to FY 15 national leads, are put into the test set. Finally, we remove ZIP codes not in the continental United States, accounting for approximately 0.6% of all national leads and use only those ZIP codes that have at least one residential address.

This research compares two different estimates of population size, United States Census Bureau county estimated population size mapped to the ZIP code level and Internal Revenue Service (IRS) ZIP code estimated population size. To do this, we utilize a United States Department of Housing and Urban Development data set to map county-level Census estimated population size data to the ZIP code level. We also utilized IRS estimated population size at the ZIP code level. We find that Census and IRS estimated population sizes have a 0.66 correlation. We find that IRS estimated population size accounts for a large proportion of the variability in the number of leads, 83%, whereas Census estimated population size accounts for only 39%. Therefore, we utilize IRS estimated population size for our models.

We develop statistical models to determine the key drivers of the number of leads at the ZIP code level. This research uses number of Qualified Military Available (QMA)

by Potential Rating Index for Zone Improvement Plan Code Markets New Evolution (PRIZM NE) segments and variables based on clustering similar ZIP codes using five sets of publicly available variables to gain insight into each ZIP code. Using Poisson regression models, we model the number of leads using number of QMA by PRIZM NE segment, five cluster membership factors constructed from publicly available sources, and IRS estimated population size at the ZIP code level.

We show that a Poisson regression model, using IRS estimated population sizes and five ZIP code cluster membership variables based on publicly available data accounts for 92% of the model deviance. Adding the 66 variables corresponding to the number of QMA by PRIZM NE segments may improve the model fit some, but makes it much more difficult for NRC to interpret and use. We recommend that NRC utilize the Poisson regression model in order to determine high-yield ZIP codes. Further, we recommend that IRS estimated population size be utilized to augment or replace the current population size variables.

ACKNOWLEDGMENTS

Thank you to my parents, Dorothy and Mike, and to the rest of my family: you have made this possible. Carol, thanks for picking up the phone when I called in the middle of the night. Grandpa, thanks for driving thousands of miles with me.

To Dr. Whitaker, Lieutenant Colonel Alt, and Dr. Buttrey: thank you for providing your expertise and helping me to keep my nose to the grindstone.

To my Monterey friends: you have made Monterey amazing and I will truly miss being able to see you every day. Fair winds and following seas…hopefully we will meet again soon.

To the “Wolfpack”: wow…where to start. Each adventure has been more and more amazing.

Last but not least, thank you to the Fireteam. Over the past two years your support has kept me going on late nights, Saturday study sessions, and Sunday Fireteam dinners. This would not have been possible without you. I cannot say thank you enough.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. UNITED STATES NAVY RECRUITING COMMAND

The United States Navy Recruiting Command (NRC) determines the number of recruits required to keep the Navy staffed, while attempting to control the costs associated with these accessions. NRC consists of two Navy Recruiting Regions, 26 Navy Recruiting Districts, and a multitude of Navy recruiting stations throughout the world (NRC 2009). In the United States, NRC assigns each Zone Improvement Plan (ZIP) code to a Navy recruiting station (NRC 2009). While Navy recruiting stations may cross state and county boundaries, they do not cross ZIP code boundaries.

NRC uses a Recruiter Allocation Factor (RAF) to determine the number of recruiters assigned to a Navy recruiting station (NRC 2009). NRC calculates the “traditional” RAF using the three previous years of accessions for 17–21-year-old males going on active duty per ZIP code (NRC, 2009). NRC then determines a recruiter’s workload, or the recruiting goal, based on the ratio of the number of recruiters to the number of qualified military available (QMA) per ZIP code (NRC 2009). It is well known that the number of accessions is strongly related to the number of recruiters assigned to a region (Pinelis et al. 2011). Thus, the number of accessions is not a good measure of market depth or the number of potential recruits in a region. In this research, we explore an alternate measure of recruiting potential and we explore this measure at the ZIP code level.

We define a national-level lead as a lead generated on the Navy.com website or through social media. We define a local lead as lead generated either at a local event held by a recruiter or by an individual going to the local recruiting station. Since the number of local leads is likely to be related to regional recruiting efforts and local recruiter’s efforts, our focus is on national leads. While an individual may be interested in joining the military, that individual may not be qualified or eligible to join for a number of reasons. A 2014 article by the Wall Street Journal highlights that “more than two-thirds of American Youth Wouldn’t Qualify for Service” (Jordan 2014). According to the article, some of the reasons an individual may not qualify for the military include obesity, aptitude, and drug use (Jordan 2014). Recruiting qualified and eligible individuals presents a continuous challenge to the military recruiting enterprise and consumes a large amount of resources. “From 2000 to 2008, the Defense budget for enlistment bonuses more than doubled to $625 million” (Jordan 2014). In a 2003 report, the Government Accountability Office (GAO) highlights that the U.S. Department of Defense’s (DOD) total recruiting budget reached $4 billion annually (U.S. GAO 2003).

This research provides NRC leadership with insight into the depth of the recruiting market at the ZIP code level in order to aid their resource allocation decisions. ZIP codes yield recruits at different rates depending on a variety of factors (e.g., socio-economic, population and propensity) so by identifying ZIP codes with a higher potential for producing leads, recruiting leadership can align their recruiting efforts to an area’s potential. This analysis augments NRC analyses, which use the number of prior years’ accessions in an area to inform its knowledge of the depth of market in a geographic region.

B. PROBLEM STATEMENT

This research seeks to address the problem of determining a geographic area’s potential recruiting market without relying on previous performance, the number of recruiters assigned to the area, or the area’s QMA. We utilize the number of national leads as a surrogate for an area’s market depth. We further decompose the problem into the following questions:

1. What factors influence the number of leads?
2. Can we predict the number of leads using publicly available data?

C. SCOPE OF RESEARCH

This research develops statistical models of the number of ZIP code-level national leads. As we will discuss in Chapter III, we use variables constructed from publicly available data that represents economic, demographic, health, education, and military regional characteristics. Our research will only utilize ZIP codes within the Continental United States (CONUS).

1. Constraints

At the request of the sponsor, we exclude the number of accessions from our research. Previous research demonstrates that historic numbers of accessions and recruiting effort is strongly correlated to future accessions (Pinelis et al. 2011, Intrater 2015, Marmion 2015). By using the number of leads as our response variable, we attempt to gain insight into the market depth while minimizing the influence of historic recruit production.

Second, we also exclude the number of QMA in a ZIP code from our models. The number of QMA is well-known and used often in modeling accessions (Pinelis et al. 2011, Intrater 2015, Marmion 2015, Parker 2015). Developed by Woods & Poole Economics, the number of QMA uses demographics, employment, health, crime, income, retail, average ASVAB score, and education data to estimate the number in the population available for recruitment at the ZIP code level (Woods and Poole, 2013).

Third, at the request of the sponsor, we exclude the number of recruiters assigned to a station. Previous research conducted by Gibson et al. (2011) shows that as recruiters are added to an area, Army accessions increase. Building upon this work, Intrater (2015) shows a strongly positive linear relationship between the number of Navy accessions and the number of recruiters.

2. Limitations

With the exception of the number of leads, we limit the scope of this study to data that is publicly available in a manner similar to Fulton (2016). We also limit this research to leads with an interest in enlisting in the active duty Navy. Like Fulton, who studies similar issues for United States Army Recruiting Command (USAREC) but with different emphasis, we limit the scope of this research to CONUS ZIP codes, excluding data from Alaska, Hawaii, or territories. As Fulton (2016) points out, the availability of publically available health, economic, and other data for some of these regions is far less than for CONUS. Additionally, in a similar manner to Fulton (2016) we reduce the number of ZIP codes we use because we only use ZIP codes with residential addresses, typically removing Post Office (P.O.) Boxes or ZIP codes zoned only for businesses. Chapter III contains a more detailed discussion of the data.

3. Assumptions

Much of the data is collected at the county level and must be mapped to the ZIP code level for our purposes. We assume that our mapping is reasonable and that ZIP codes correspond to the same recruiting station every year. We know that ZIP codes can span county and state lines and can change annually according to United States Postal Service (USPS) needs (USPS 2015). As our aggregated leads data set covers four years, following Intrater's (2015) methodology, we treat ZIP codes as if they stay the same year to year. Chapter III contains a more detailed discussion on mapping county level data to the ZIP code level.

Finally, we assume that the number of leads in the Navy data set does not contain a large number of duplicate records. As Darrow (2016) notes, the USAREC leads utilized in support of his research contain duplicate records, with duplicates occurring when a lead occurs at both the national and local level. As we do not have access to the database of leads that the Navy maintains, we do not know if the data set contains duplicate records. We assume that as we remove local leads in order to remove recruiter influence, we mitigate the possibility of duplicate leads from our model.

D. RESEARCH OUTLINE

We organize the document as follows: Chapter II covers previous research on estimating military accessions. Chapter III provides a description of the data sets and data preparation. Chapter IV focuses on the analysis of the data and models created during this research. Chapter V provides recommendations for which model to utilize and suggested future research.

II. BACKGROUND

A. RECENT RESEARCH

In a recent investigation, Williams (2014) examines factors that could influence Navy recruiting. Williams analyzed the Noble Index, an index developed by NRC to gain insight into the market potential of a geographic area. He fit both long and short-term models to predict recruiting potential using accessions and found that for the long-term, a linear regression model worked best. In the monthly model, Williams found the Poisson regression model to be best at predicting accessions. The two most important factors he identified for both models were (1) number of Navy recruiters per geographic area and (2) the "competition index" or the number of other services competing within the same geographic area for the same enlisted accessions (Williams 2014). Williams found that as the number of Navy recruiters increased, the number of accessions increased; he also found that the higher the competition index value the smaller the number of accessions. At the end of his research, Williams recommends using leads as a proxy for recruiting potential.

Intrater (2015) examined how socio-economic factors impacted enlisted accessions. Consistent with Williams, at the station-level, Intrater found the number of recruiters assigned to an area to be the most significant factor impacting the number of enlisted accessions. At the ZIP code level, he found "Recruiter strength, total veterans, underprivileged [areas defined as areas where adjusted gross income is less than $25,000], and violent crime are the strongest positive predictors" (Intrater 2015). Since 2015, the Navy has been using a multivariate regression model to predict what accessions a recruiting district should be able to attain in the near future based on manpower and economic trends (Ammons-Moreno 2016).

B. MARKET SEGMENTATION

Neilsen, previously known as Claritas, provides a customer segmentation system called Potential Rating Index for ZIP Code Markets New Evolution (PRIZM NE) (Neilsen 2015). Neilsen's segmentation models provide insight into the type of individual most likely to use a product and how to tailor advertising to them (Neilsen 2015). Using demographics, geography, survey media usage, and survey leisure data, Neilsen groups a household into segments per ZIP code based primarily upon income, education, spending habits, home value, age, and children.

United States Army Recruiting Command (USAREC) uses the Army Custom Segments (ACS), built by Integras in 2005, to understand American youth (Moffit 2016). Integras surveyed the youth population on military enlistment and combined the survey results with PRIZM NE segments, going from 66 PRIZM NE segments to less than 40 ACS (Moffit 2016). As a recruiter learns race, ethnicity, gender, and date of birth of a lead, that lead gets binned into an ACS; in the early stages of recruitment, without this information, a lead cannot be binned (Moffit 2016). Marmion (2015) finds as more partitions are used in modeling accessions, a market's potential accessions can be more accurately modeled. Currently, NRC does not use PRIZM NE data in its accession models.

C. ZIP CODE LEVEL MODELS

Pinelis, Schmitz, Miller, and Rebhan (2011) developed a zero-inflated Poisson model at the ZIP code level. The authors used demographics, distance to the responsible Navy recruiting station, a self-created Navy Awareness Index, the number of recruiters in a geographic area, crime data, and veteran population as predictor variables (Pinelis et al. 2011). The model uses five years of historic accession data to predict the next year's accessions. They found in their analysis that veterans age 45 to 84 negatively influenced accessions. Pinelis et al. further concluded, as explained by Intrater, "Navy awareness and crime data all had positive association with accessions, but distance to the responsible Navy recruiting station had a negative association" (Intrater 2015).

Gibson, Hermida, Luchman, Griepentrog, and Marsh (2011) utilized a zero-inflated Poisson regression model in their paper entitled "The Armed Services ZIP Code Valuation Study," building off a model previously developed in 2009. The authors found that a significant number of ZIP codes never yield Army recruits, which holds for Navy leads as well (Gibson et al. 2011). The authors used predictor variables gathered at state and county levels; in a manner that Fulton (2016) mimics, Gibson et al. (2011) mapped data available only at the state or county level to the ZIP code level using distances. The authors then modeled accessions per ZIP code.

The authors found that as the number of recruiters per geographic area increased, the number of enlisted accessions increased; United States Marine Corps recruiters were the greatest inhibitors to other services’ accessions (Gibson et al. 2011). Gibson et al. also found that as American College Test scores in a ZIP code increased, a ZIP code would be less likely produce enlisted recruits.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA COLLECTION AND PREPARATION

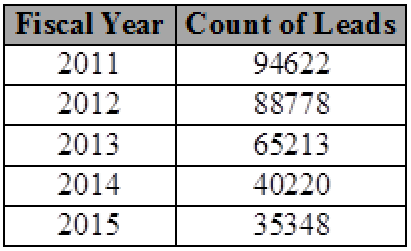
In this chapter, we discuss the data sources and some of the preparation required to use data from these sources. We also mention the modeling techniques used in this research.

A. DATA COLLECTION

1. National Leads

NRC collected and provided records of Navy recruiting leads generated from Fiscal Year (FY) 11 to FY 15 (NRC 2016a). The data set contains approximately 620,000 observations and for each date, and each ZIP code contains the number of leads generated by type of lead (NRC 2016a). Type of lead, in this case, refers whether the lead is a national or a local one. In order to remove any possible influence recruiters might have upon leads, given the strong association between accessions and recruiters shown by Gibson et al. (2011) and Intrater (2015), we remove local leads, which account for approximately 7% of all leads. Table 1 shows the count of national leads per year. As we can see from the counts, the number of leads generated each year has been sharply decreasing since FY 11.

Table 1. Number of National Leads per Fiscal Year



Fiscal Year	Count of Leads
2011	94622
2012	88778
2013	65213
2014	40220
2015	35348

Figure 1 is a boxplot of the logarithm of the number of leads per ZIP code grouped by fiscal year. The x-axis gives each fiscal year in the data set. The y-axis gives the number of leads per ZIP code transformed on the logarithmic scale. Throughout, when taking the logarithm of the number of leads, the number of leads is assigned a value of 0.05 for those ZIP codes with no leads. We can see that there is a substantial difference between the distribution of leads from FY 11 and those from FY 15.

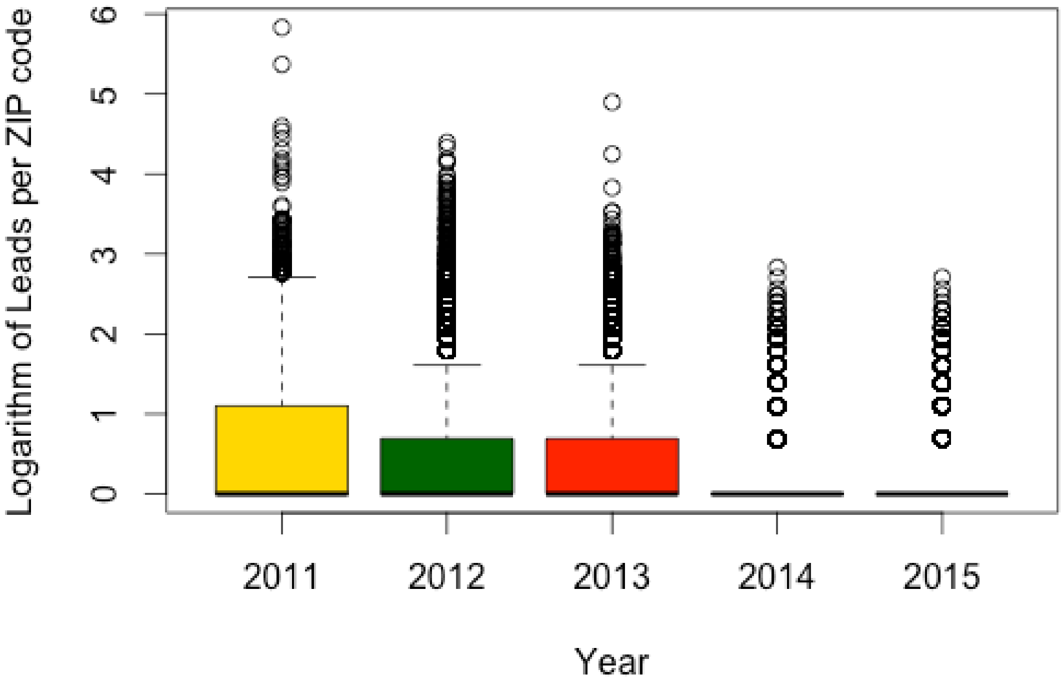
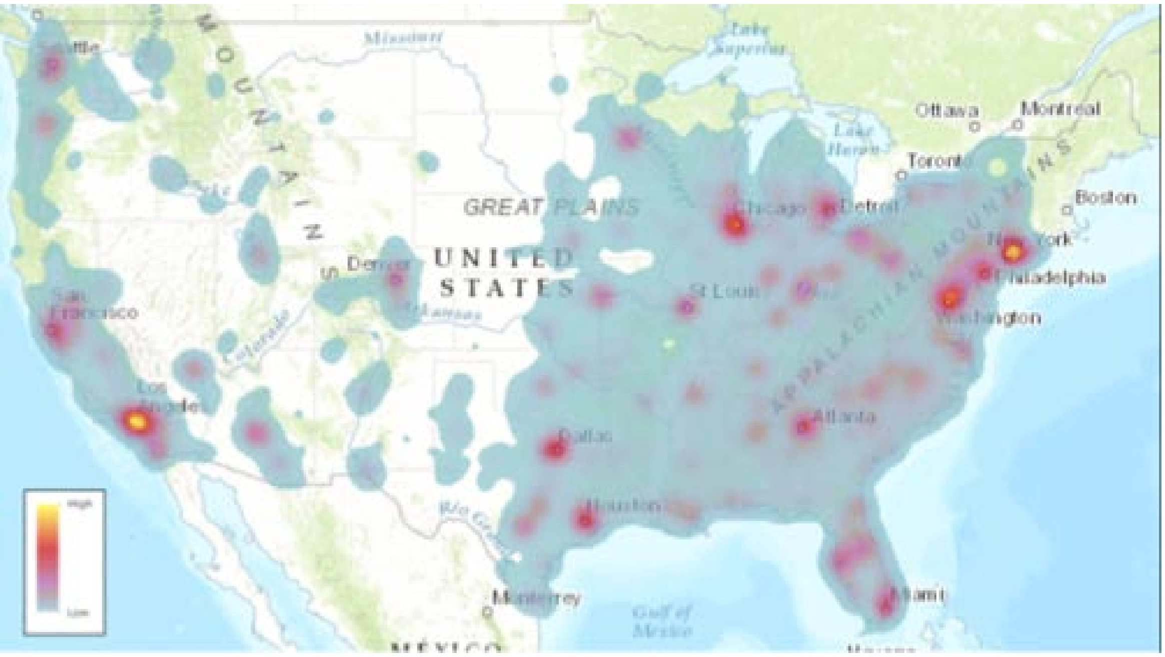


Figure 1. Boxplot of the Logarithm of Leads per Fiscal Year

Table 1 and Figure 1 both show a continual decline in the number of leads. FY 15 represents only 10.9% of the number of leads for the full five years. This is about half of what we expect, possibly indicating a change in the data collection process or some other change affecting interest in the Navy. Every year NRC either meets or minimally deviates from its stated recruitment goal (NRC 2016b). On the other hand, in recent years, the Navy's end strength has taken a "deeper-than-expected cut" forcing the Navy to downsize (Larter and Faram 2016).

Next, we aggregate national leads to create a ZIP code-level model. The training set utilizes national leads data from FY 11 to FY 14. We sum the leads over four fiscal years by ZIP code to generate the four-year total number of leads for the training set for each ZIP code. The number of leads from FY 15 is set aside as a test set. Finally, we remove ZIP codes not in the continental United States (CONUS), approximately 0.6% of all national leads.

Figure 2 provides a heat map of the number of national leads from FY 11 to FY 14, showing a large number of leads on the east coast. The figure is on the scale of the logarithmic transformation of the number of leads where yellow corresponds to highest frequencies of leads, followed by red and then blue-gray areas. We note that the heat map in Figure 2 is based on ZIP code centroids rather than ZIP code boundaries so as not to inflate the importance of ZIP codes that cover large areas.



The number of national leads between FY 11 and FY 14 on the logarithmic scale based on ZIP code centroids.

Figure 2. Heat Map of the Number of National Leads from FY 11 to FY 14

2. Number of Qualified Military Available by Potential Rating Index for ZIP Code Markets New Evolution Segment Data

As discussed in Chapter II, the PRIZM NE data set comes from surveys conducted by the Neilsen Marketing Research group. According to the Segmentation Marketing Guide, PRIZM NE's purpose is to "provide users with military-relevant information about the attitudes and interests of youth" (Neilsen 2015). The 66 PRIZM NE segments are designed to provide a better understanding of each ZIP code so that Neilson's customers can better tailor their recruiting and marketing activities (Neilsen 2015). USAREC provides this data set, which gives number of QMA by PRIZM NE segment for each ZIP code (USAREC, 2016). Currently, NRC does not use number of QMA by PRIZM NE segmentation data.

3. Housing and Urban Development Data

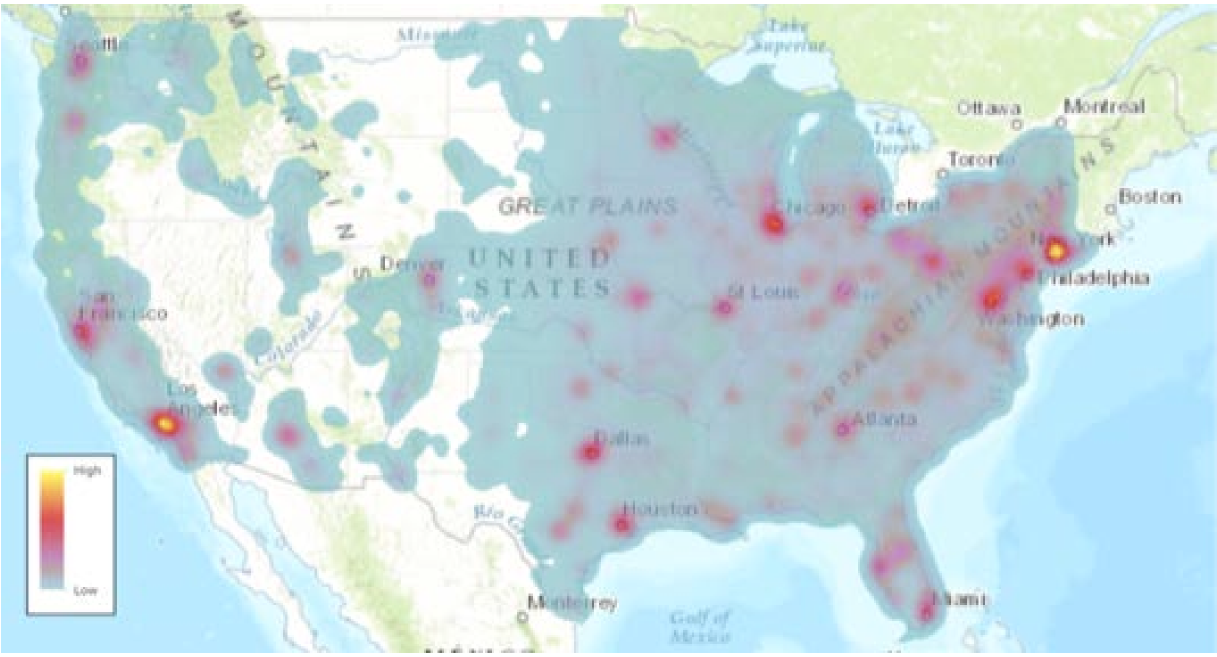
United States Department of Housing and Urban Development publishes the HUD ZIP Code crosswalk data set every three months at the county and state levels (HUD 2016). HUD utilizes USPS vacancy data, also published every three months, to determine the number of residential addresses that are occupied, referred to as residential ratio, and number of businesses in a ZIP code (HUD 2016). Approximately 83% of all ZIP codes contain residential addresses. ZIP codes that typically do not have residential addresses are P.O. Boxes or ZIP codes zoned only for businesses. In a manner similar to Fulton (2016), we use only ZIP codes that have one or more residential addresses. This reduces the total number of ZIP codes by approximately 5,300 ZIP codes. We use the HUD data set to map county level data to the ZIP code level. While there are geographic areas where population undergoes changes or migration, we assume these will not affect the overall classification process (HUD 2016).

4. United States Census Bureau Estimated Population Data

This study explores two different estimates of population size. The first estimate of population size we explore is the U.S. Census. The U.S. Census Bureau (Census) collects population data every ten years and the most recent is from 2010. We did an initial analysis on both the 2010 base estimate of population size and the estimated population size for 2014. Base estimate of population size refers to the population count used as the starting point in the estimation process done by the Census, while estimated population size refers to the Census' best prediction on the current population (U.S. Census 2016). We choose to utilize the 2014 estimate of population size, as it is the most current estimate of population size produced by the Census.

Next, we extrapolate the 2014 Census estimated population size from the county-level to the ZIP code-level. The Census gathers population data at the ZIP code tabulation area (ZCTA) level and county-level (U.S. Census 2016). We use the county-level data to simplify the translation to ZIP code-level, utilizing the HUD data set to translate.

During our exploratory analysis, we use ArcGIS to map the Census estimated population size data (ESRI, 2016). The heat map of the estimated population size for the United States, shown in Figure 3, shows the highest area of population centers on the greater Los Angeles area with the general Boston to New York City area following closely behind. We use the logarithmic transformation of the estimated population size in order to improve visualization for the heat map. Yellow corresponds to highest population estimates; red then blue-gray areas on the map also correspond to middle and small population densities respectively. From Figures 2 and 3, we see that the number of leads generated is closely related to population size.



Heat map of the estimated population size for 2014 by the U.S. Census Bureau transformed on the logarithmic scale.

Figure 3. Heat Map of U.S. Census Estimated Population Size for 2014

5. Internal Revenue Service Estimated Population Data

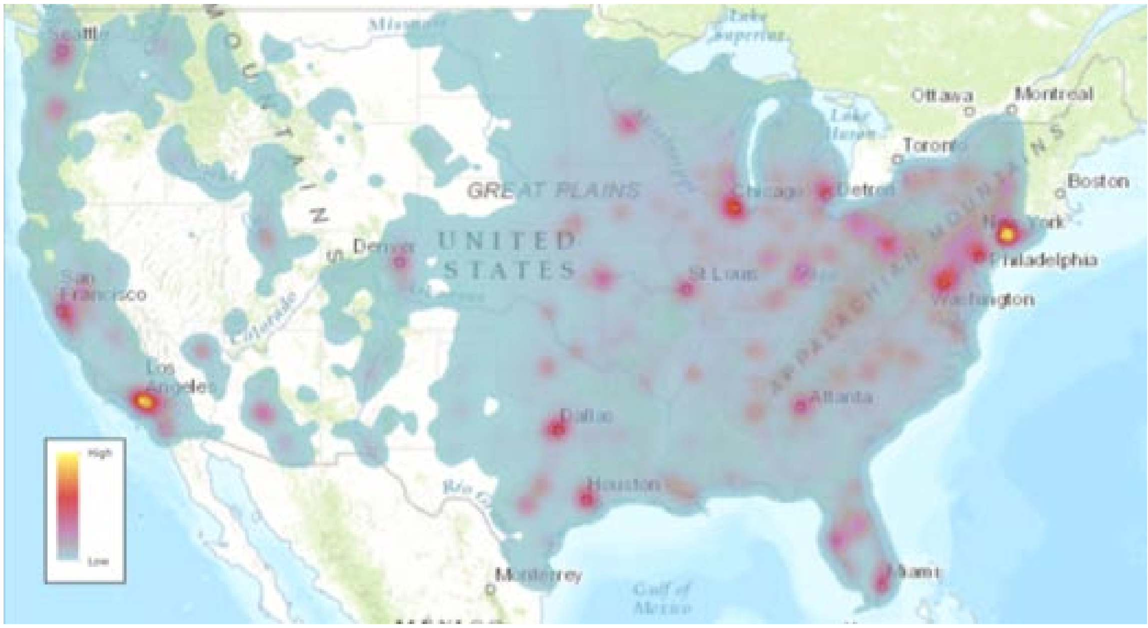
The second estimate of population size we explore comes from the Internal Revenue Service (IRS). The IRS collects taxes from the citizens of the U.S. every year. Each year the IRS publishes ZIP code level data based on its administrative records of individual income tax returns, Form 1040 (IRS 2016). We use data based on individual income tax returns filed from January 1 2014 to December 31 2014 (IRS, 2016). It is important to note that the “data does not represent the full U.S. population because many individuals are not required to file an individual tax return” (IRS 2015). Further, the IRS excludes tax returns filed without a ZIP code and returns filed where the ZIP code does not match the state on the return (IRS 2015).

The IRS data set only includes ZIP codes with more than 20 filed returns, rounding all counts such as the number of returns filed in the data set to the nearest 10 (IRS 2015). Per IRS documentation, the IRS turns “ZIP codes with less than 100 filed returns and those identified as a single building or nonresidential ZIP code” into ‘99999’

ZIP codes, rounding to the nearest 10 (IRS 2015). Our review of the raw IRS data set shows that while the IRS does round to the nearest 10, there are instances of 90 returns filed, leading us to believe that the cut off for returns is actually 90 rather than 100.

For this research, we use number of exemptions per ZIP code as our estimate of population size. Exemptions count each taxpayer and each dependent of the taxpayer (IRS 2015). A dependent must be related to the taxpayer, live in the same domicile, and not file his or her own tax return (IRS 2015). Children qualify as dependents if under the age of 24 and going to school full time or if permanently disabled (IRS 2015). For ZIP codes containing or near educational institutions such as universities, the student population will not be counted towards that ZIP code, but towards their parent's ZIP code.

In a 1998 article, Sailer and Weber (1998) use the IRS number of exemptions to estimate population size, comparing their estimates with Census population size estimates. The authors first clean the data using social security numbers, excluding foreign addresses, and accounting for individuals who did not file a tax return. Sailer and Weber (1998) find that they can estimate the population size to within approximately 96% of the Census population estimate. Sailer and Weber (1998) had extensive use of the internal database at the IRS and make use of data cleaning techniques that are not available for this research. We use the raw (rounded) number of exemptions as the IRS estimate of population size. We can see that our heat map of the IRS estimate of population size, shown in Figure 4, appears to approximately match the Census estimate of population size. Further comparisons are made in the next chapter.



Heat map of the Internal Revenue Service logarithm of estimated population size for 2014.

Figure 4. Heat Map of IRS Estimated Population Size for 2014

6. ZIP Code Cluster Memberships

In this research, we use five cluster membership factors at the ZIP code-level constructed by Fulton (2016) and used by him to predict U.S. Army national leads. This data is unpublished but Fulton (2016) describes its construction in enough detail that it can be reconstructed. Fulton’s five cluster membership factors are based on variables taken directly from publicly available sources or are constructed from publicly available sources. Fulton uses 347 variables and separates them into five categories: economic, military, health, education, and demographic. The sources of raw data “include the Internal Revenue Service, the Census Bureau, the U.S. Post Office, the Centers for Disease Control and Prevention, and the Department of Housing and Urban Development” (Fulton 2016). For example, a few of the variables in the demographic category of variables that Fulton uses are: the population size from the Community Health Status Indicators (CHSI), percentage of population age from CHSI, and percentage of filed tax returns filed single from the IRS. Further, it is important to note that Fulton partitions the 347 variables into five separate categories without overlap and that variables from the same source can be categorized differently. For example, Fulton separates the variables extracted from the CHSI data set into four out of the five cluster membership factors, but does not allow the variables to be used twice. A full listing of the variables used, their source, and the description of each variable can be found in Appendix A of Fulton's (2016) thesis.

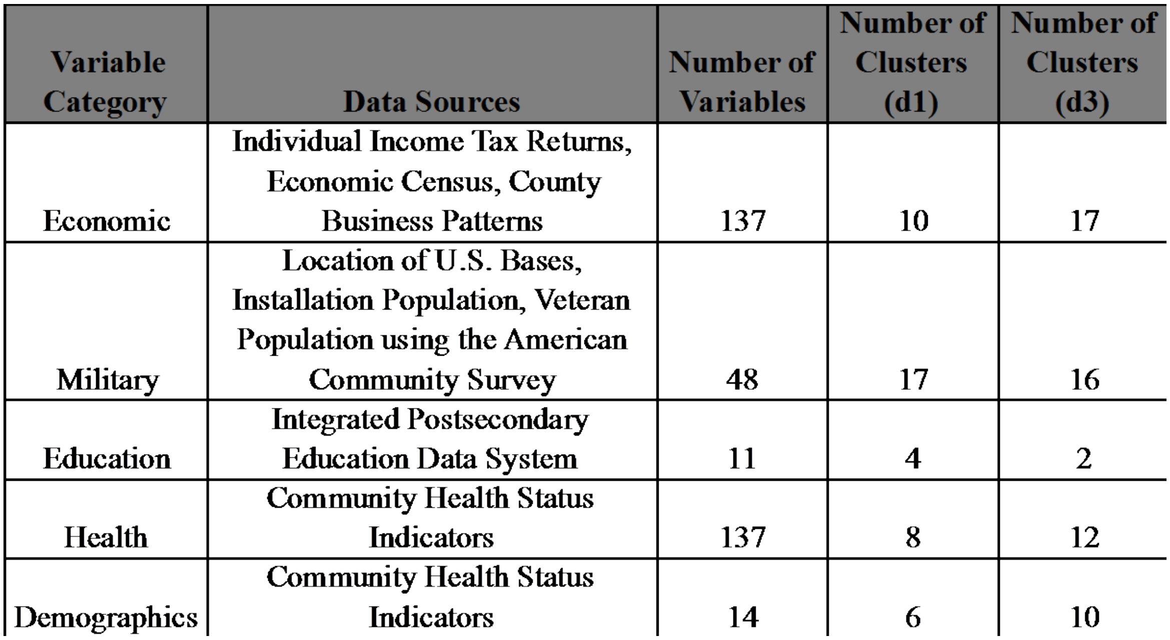
In his data cleaning and preparation, Fulton (2016) maps several variables to the ZIP code-level in a manner similar to that of Gibson et al. (2011) and Pinelis et al. (2011). Fulton also utilizes the HUD data set to map variables to the ZIP code-level (Fulton 2016). Finally, Fulton (2016) changes variables that are counts per person into percentages of the total estimated population size. To compute this percentage, Fulton utilizes the CHSI estimated population size for the CHSI data and the Census population size for all other data sets. At the end of data preparation, Fulton's data is all at the ZIP code-level.

Fulton (2016) forms five cluster membership factors at the ZIP code-level based on each of the five sets of variables. The five cluster membership factors indicating ZIP code cluster memberships then replace the five sets of variables. Replacing the original 347 highly linearly dependent variables constructed from publicly available sources with five cluster membership factors at the ZIP code level gives a set of variables that NRC can more easily use and interpret.

Fulton (2016) clusters each of the five sets of variables using the *treeClust* function in R (Buttrey and Whitaker 2015). The *treeClust* function produces distances or dissimilarities between observations that allow for missing values, of which there are many in the raw data; they also allow for both numeric and categorical variables and are invariant to scale changes of the numeric variables (Buttrey and Whitaker 2015). The *treeClust* function returns one of four versions of dissimilarity; see Buttrey and Whitaker (2015) for a detailed description of the dissimilarities and their comparison with other methods.

Fulton (2016) finds that two of the dissimilarities, *d1* and *d3*, in combination with the clustering methods Partitioning around Medoids (Pam) and k-means, perform similarly when used to predict the number of Army national leads per ZIP code. For this research, we compare clusters produced by Pam based on *d1* and *d3.* The number of clusters based on the five categories of variables varies from two to eighteen depending on the category and the dissimilarity, *d1* and *d3*, used. Table 2 shows the number of variables in each category of variables and the number of clusters used by Fulton based on *d1* and *d3* dissimilarities.

Table 2. Categories of Clusters for *d1* and *d3* Pam



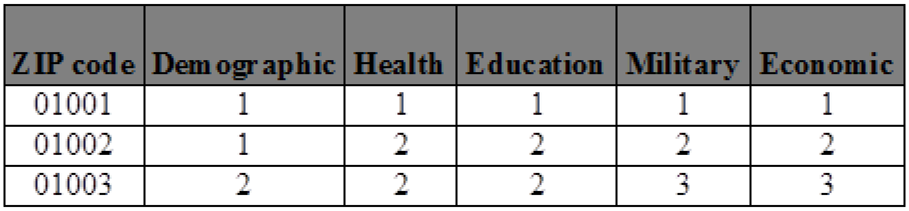
Variable Category	Data Sources	Number of Variables	Number of Clusters (d1)	Number of Clusters (d3)
Economic	Individual Income Tax Returns, Economic Census, County Business Patterns	137	10	17
Military	Location of U.S. Bases, Installation Population, Veteran Population using the American Community Survey	48	17	16
Education	Integrated Postsecondary Education Data System	11	4	2
Health	Community Health Status Indicators	137	8	12
Demographics	Community Health Status Indicators	14	6	10

This table shows each of the variable categories, the data source for each category, the number of variables for each category, and the optimal number of clusters for each category (Fulton, 2016).

We note that there is no one “optimal” number of clusters to use and there are a number of methods that can be used to identify a reasonable number of clusters that might be used in any situation (Maechler et al. 2015). We use the number of clusters found by Fulton (2016) since they perform well in predicting the numbers of Army national leads by ZIP code.

An example of the *d3* Pam cluster memberships per ZIP code can be seen in Table 3. Here ZIP codes 01001 and 01002 fall in cluster 1 of the 10 clusters of ZIP codes formed based on the 14 demographic variables, whereas ZIP code 01003 falls in demographic cluster 2. We note that the cluster membership variables are categorical and that the numbers in Table 3 only indicate to which cluster each ZIP code belongs.

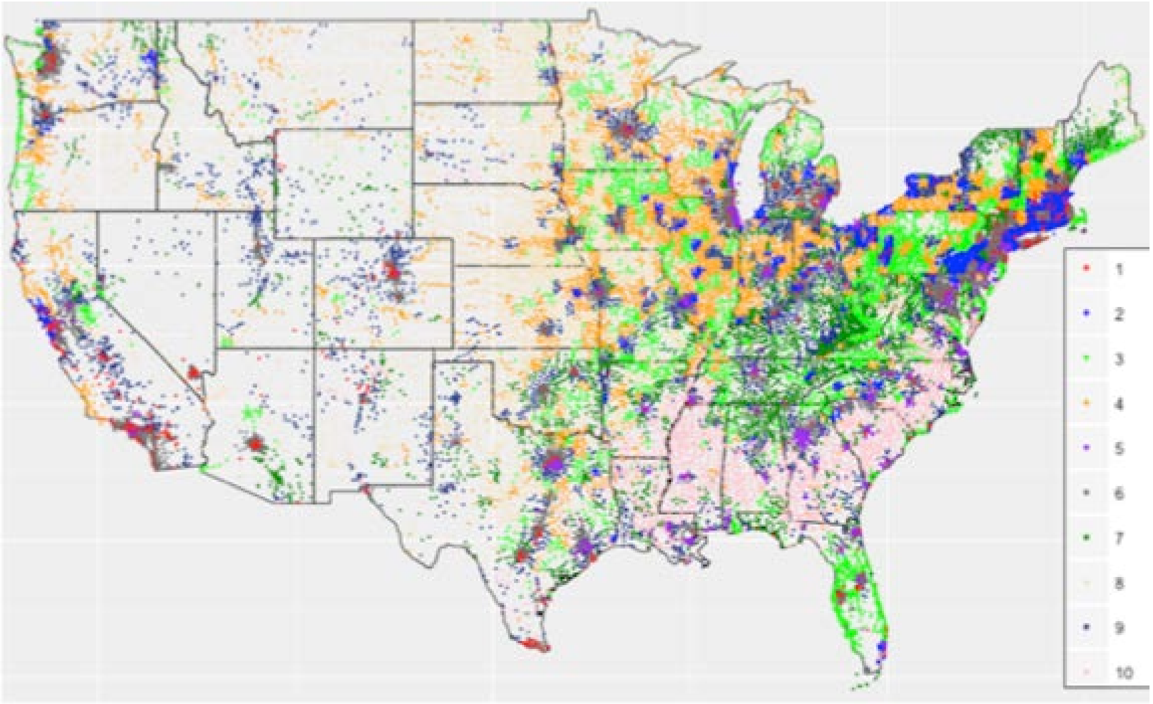
Table 3. Cluster Memberships



ZIP code	Demographic	Health	Education	Military	Economic
01001	1	1	1	1	1
01002	1	2	2	2	2
01003	2	2	2	3	3

This table is an example of the cluster memberships at the ZIP code level provided by MAJ Fulton (2016). This table shows clusters constructed using *d3* Pam.

We map ZIP codes by *d3* Pam cluster memberships obtained from Fulton (2016) to gain a better understanding of the associated clusters. Maps for all categories can be found in Appendix C, as well as discussion of these maps. Figure 5 shows the ZIP code cluster membership based on the demographic variables. The dots in Figure 5 represent ZIP code centroids, colored according to membership in one of the ten demographic clusters.



treeClust cluster assignments using *d3* Pam obtained from Fulton (2016). The legend in the bottom right corner shows each of the colors associated for each cluster.

Figure 5. Map Showing ZIP Code Cluster Assignments Based on Demographic Variables.

Cluster 1 is red; clusters 2 through 10 are depicted with blue, green, orange, purple, gray, dark green, wheat, royal blue, and pink respectively.

The gray cluster corresponding to cluster 6 has the highest number of ZIP codes of all clusters, 4,786. This cluster is found primarily in California (with 771 ZIP codes) and Texas (with 513 ZIP codes). Cluster 6 can be characterized as having a high average population count, approximately 11,900; a high percentage of that population under 19 (27.6%); the highest average Asian population of all clusters (5.4%); and a very high Hispanic population (16%).

4,566 ZIP codes are the orange color, corresponding with cluster 4. Cluster 4 is found primarily in Illinois (with 442 ZIP codes) and New York (with 338 ZIP codes). Cluster 4 ZIP codes can be characterized as smaller average population size with approximately 3,800 people but a very high average white population (94.1%).

The bright green color corresponds with cluster 3, consisting of 4,464 ZIP codes. This cluster is typically found in Pennsylvania (with 617 ZIP codes) and Florida (with 605 ZIP codes). Cluster 3 ZIP codes can be characterized as a mid-size, with an average population size of 4,900, a high percentage of the population between the ages of 65 and 84 (16.1%), and a high average white population (93%).

The bright yellow cluster corresponds to cluster 8. Cluster 8 does not contain a large number of ZIP codes, with 2,907 total ZIP codes, but can be seen in tight groupings throughout the northeastern part of the United States. This cluster is very interesting as it has the highest average percentage of all clusters for population between the ages of 65 and 84 (16.8%), the highest average percentage of all clusters for average population older than 85 (3.37%), the highest percentage of all clusters for average white population (96%), the smallest average percentage of all clusters for black population (1.06%), and the smallest average percentage of all cluster for Asian population (0.39%).

Finally, we can easily see the pink color throughout the southern portion of the United States, corresponding to cluster 10. Cluster 10 only consists of 2,000 ZIP codes, but all of those ZIP codes correspond to the same general geographic area of the United States, making it stand out. Cluster 10 ZIP codes are primarily found in Mississippi, Louisiana, Alabama, Georgia, North Carolina, and South Carolina. This cluster can be characterized as having a high average population size, approximately 7,900, a high average population under the age of 19 (26.7%), the highest average population of all clusters for black population (34.5%), and the highest average population of all clusters for Native American population (2.7%).

B. DATA PREPARATION

First, we examine the NRC leads data set to identify missing values or other data anomalies. Following the collection and initial cleaning of the individual data sets, we construct the final data set.

We start by removing local leads from our data. We then separate leads into training and test sets. The training set takes leads data from FY 11 to FY 14; leads from FY 15 are put into the test set. Then, we sum the number of leads by year for each ZIP code to generate a single value of the number of leads for the training set for each ZIP code. For ZIP codes that never generate any leads, we change the number of leads from zero to 0.05 in order to use the logarithmic transformation of the number of leads in our analysis.

Next, we import the Census estimated population size data. As Census estimated population size does not have associated ZIP codes, we generate each county's associated Federal Information Processing Standard (FIPS) county code. We then, in a manner similar to Fulton (2016), utilize the FIPS county code database from the Census to remove any FIPS codes not in CONUS (U.S. Census, 2010) and use the HUD data set to map each FIPS code to a ZIP code (HUD, 2016). We multiply the residential ratio for each ZIP code from the HUD data set by the Census county level 2014 estimated population size, the most current available, giving us the 2014 ZIP code-level estimate of population size.

We next import the IRS data set into R. The IRS data set contains both ZIP codes and number of exemptions for each filed income tax return per ZIP code, which we are utilizing here as an estimate of a ZIP code's population size. We remove all other variables from the IRS data set because they are similar to variables included in the set of economic variables used by Fulton (2016) to construct one of the five sets of ZIP code cluster membership variables that we use.

Following our estimated population size construction, we import the number of QMA by PRIZM NE data set into R. We also import the five cluster membership factors used by Fulton (2016) at the ZIP code level. The five types of factors are economic, demographic, education, military, and health. Only one ZIP code, 70176, has no economic cluster membership. Smith et al. (2016) notes that ZIP code 70176 is a P.O. Box only ZIP code that is a part of ZIP code 70130. Using this information, we give ZIP code 70176 the same economic cluster membership value as 70130.

Finally, we construct a master data set by combining the number of leads, number of QMA by PRIZM NE segment, five cluster membership factors, IRS estimated population size, and Census estimated population size data sets at the ZIP code level. The final master data set has approximately 25,300 rows, one for each CONUS ZIP code represented in the data and 73 columns for the variables listed above as well as a few extra variables such as latitude and longitude of a ZIP code to aid in data exploration.

C. MODELING TECHNIQUES

1. Multiple Linear Regression Models

"Standard" multiple linear regression, as explained in Faraway (2015), uses a single response variable and possibly multiple predictor variables to explain the relationship between them. However, a multiple linear regression model cannot easily handle non-normal errors with non-constant variance (Faraway 2015). We use multiple linear regression as an exploratory tool. Linear regression models are fit using the *lm* function from the base *stats* package included in R (R development core team 2013).

2. Generalized Linear Models

Where the "standard" linear regression model assumes a normal response variable with constant variance, the generalized linear model can handle response variables of different types (Faraway 2006). We specifically use a Poisson generalized linear model because the number of leads is a count variable and Poisson generalized linear models have a variance structure which is often more appropriate for count variables (Faraway 2006). We fit generalized linear models using the *glm* function from the base *stats* package included in R (R development core team 2013).

3. Generalized Additive Models

Generalized additive models utilize the same basic response and predictor variables as the generalized linear model, but allow the analyst to easily estimate smooth nonparametric functions of numeric predictor variables (Faraway 2006). We fit a generalized additive model using the *gam* function from the *gam* package in R as a diagnostic tool to identify nonlinearity in numeric predictors (Hastie 2015).

4. Assessing the Model and Variable Selection

For the generalized linear model the analogue of R-squared, pseudo R-squared, is the proportion of decrease in the residual deviance from the null deviance (the residual deviance from the model fit with a constant term and no predictors) (Faraway 2006). The linear regression model fits are used only for exploration and are assessed visually using residual plots and using summary statistics. We use pseudo R-squared as a general measure of fit. We also diagnose whether numeric predictors require transformation using partial residual plots with smooth partial fits for numeric predictors. It is common for fitted Poisson generalized linear models to show evidence of over-dispersion, i.e., the variance of the response is greater than its expected value. We diagnose this by estimating the dispersion parameter with the ratio of the residual deviance to the residual degrees of freedom. When this ratio is greater than one, we fit an over-dispersed or "quasi" Poisson generalized linear model to accommodate potential over dispersion (Faraway 2006). We use a combination of large sample likelihood ratio (LRT) tests with backwards elimination for variable selection.

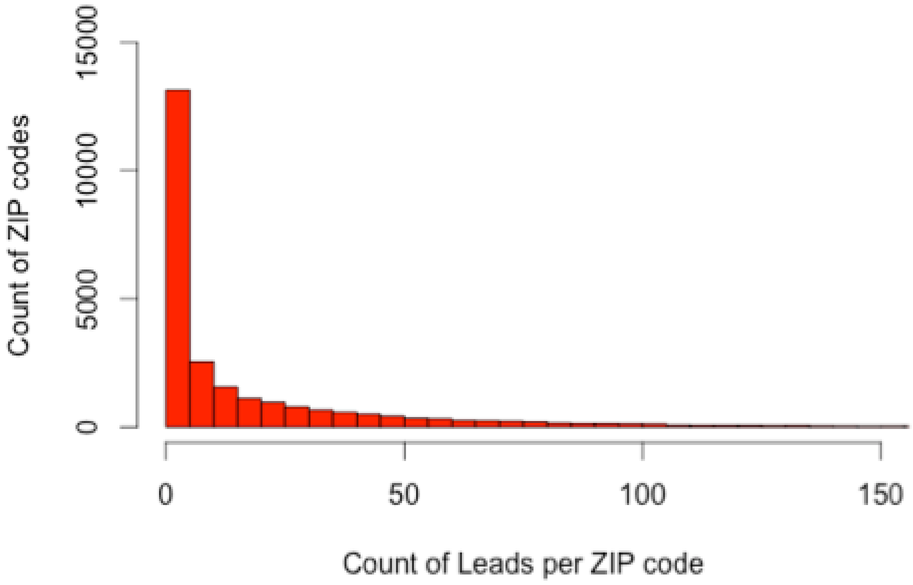
IV. ANALYSIS

A. DESCRIPTIVE STATISTICS

1. NRC Leads

Of the approximately 40,000 CONUS ZIP codes, approximately 25,300 ZIP codes are used in this study; approximately 9,400 ZIP codes are P.O. Box specific and over 5,200 ZIP codes have no residential address. 4,563 ZIP codes have no leads from FY 11 through FY 14. As we have no reason to believe that ZIP codes with no leads from FY 11 to FY 14 represent structural zeros, i.e., ZIP codes that can never produce a lead, we assign the number of leads to be zero for those ZIP codes and include them in our data set rather than deleting them.

Leads, specifically the number of leads aggregated over four FY to the ZIP code level, serve as the dependent variable in all models developed in this research. Figure 6 shows the distribution of the number of leads per ZIP code.



This histogram is of number of leads per ZIP code from FY 11 through FY 14. The x-axis is the number of leads per each ZIP code.

Figure 6. Histogram of the Number of Leads per ZIP code in the Training Set

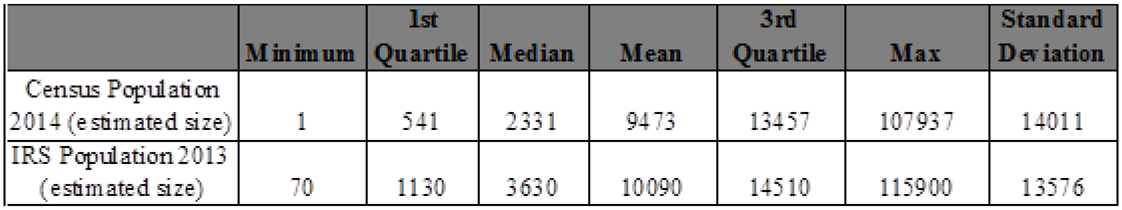
Figure 6 indicates that from FY 11 to FY 14 there is a large percentage of ZIP codes (17.7%) that did not generate leads in that four-year period; the median number of leads per ZIP code is five leads with the 25th and 75th percentiles of leads being one and 24 leads respectively. The ZIP codes with the greatest number of leads in the four-year period is 570 leads in ZIP code 30349 in Atlanta, GA.

As mentioned in Chapter III, we utilize the number of leads from FY 15 as the test set for this research. The number of leads in the test set is 10.9% of all leads from FY 11 to FY 15.

2. Population Size Variables

Table 4 shows the descriptive statistics of the two variables that we consider as estimates for population size of ZIP codes. The smallest IRS estimated population size in CONUS is 70, whereas the smallest Census estimated population size is one. We shall discuss further why the IRS estimated population size starts at 70 people.

Table 4. Descriptive Statistics of Estimated Population Size per ZIP Code



	Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Standard Deviation
Census Population 2014 (estimated size)	1	541	2331	9473	13457	107937	14011
IRS Population 2013 (estimated size)	70	1130	3630	10090	14510	115900	13576

As the distributions of population sizes are so right skewed (as indicated by the statistics in Table 4), we next compare the distribution of IRS estimated population size and the Census estimated population size on the logarithmic scale. In Figure 7 the y-axis is the logarithmic scale of estimated population size; the two different boxplots are for IRS data, in red, and Census data, in green. We note all population sizes are greater than one. As can be seen in Figure 7, the Census estimated population size distribution has several outliers and is more variable than that of the IRS.

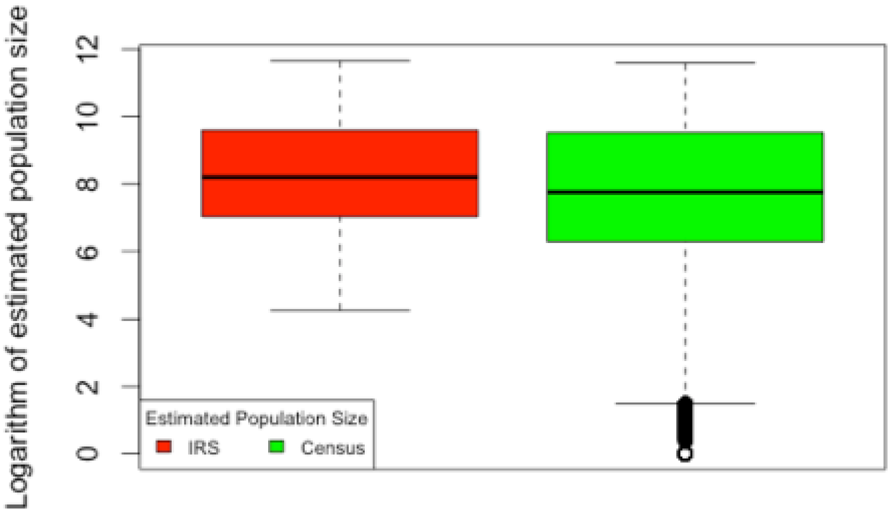
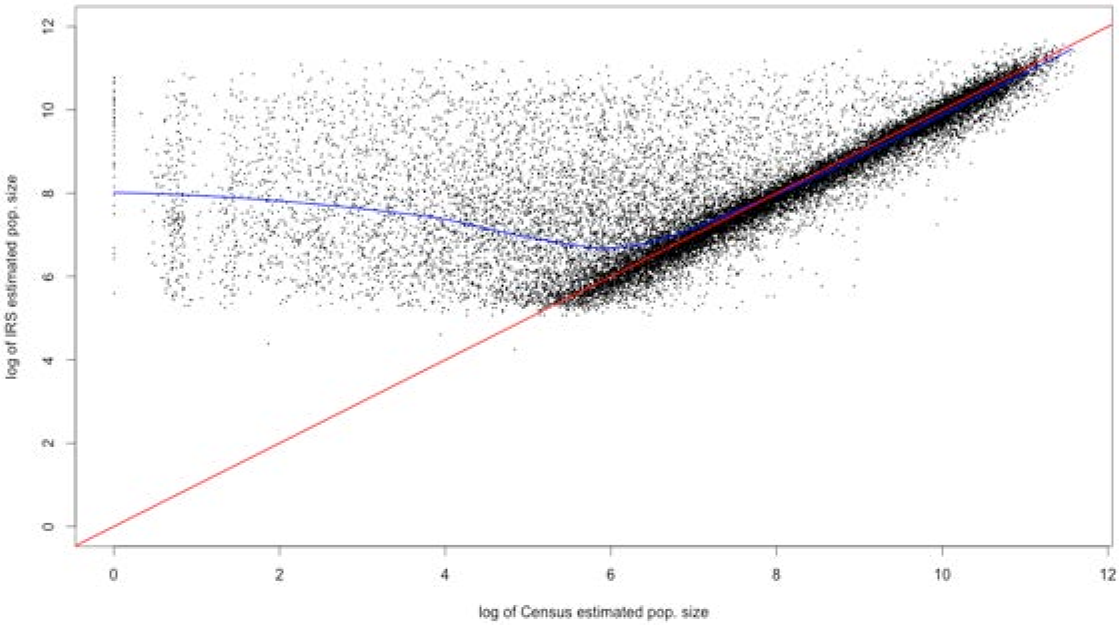


Figure 7. Boxplot of the Logarithm of Estimated Population Size

Figure 8 shows us the relationship between the logarithms of the IRS estimated population size and Census estimated population size. The correlation between the logarithms of the two population size estimates is 0.655. In Figure 8, the y-axis is the logarithmic scale of IRS estimated population size; the x-axis is the logarithmic scale of Census estimated population size; the red line has an intercept of zero and a slope of one. We can see that the IRS estimated population size cuts off around 4.2, the logarithm of 70. It is somewhat surprising that the correlation between the logarithm of the two population sizes is -0.13 when the Census population size is below 150 (corresponding to about 5.0 on the logarithmic scale). The slightly negative relationship can be seen from the nonparametric loess smoother, the blue line, in Figure 8. See Faraway (2006) for a description of the loess smoother.

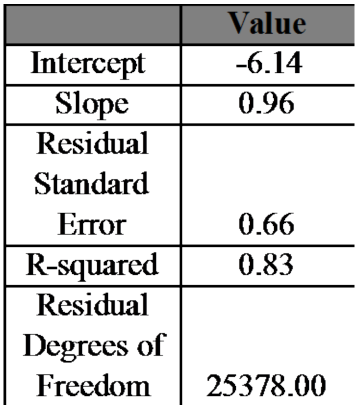


The red line has a slope of one and an intercept of zero. The blue line is the nonparametric loess smoother.

Figure 8. Scatter Plot of the Logarithms of IRS and Census Estimated Population Size

To get a sense of which estimate of population size, IRS or Census, might prove more useful in modeling the number of leads generated in a ZIP code, we fit two simple linear regression models. Both models use the logarithmic transformation of the number of leads in the training set as the response and logarithmic transformation of the estimated population size as predictors in order to determine which will provide us the best insight into the number of leads. Again, we transform both the number of leads and the estimated population sizes because both have very right skewed distributions. Note that we set the number of leads to be 0.05 for those ZIP codes with no leads. This affects only 18% of ZIP codes, as these are ZIP codes with small population sizes. The summary of the regression fit using the IRS estimated population size is given in Table 5 (standard errors and p-values of the coefficients are omitted since our linear regression model is used for exploratory purposes only).

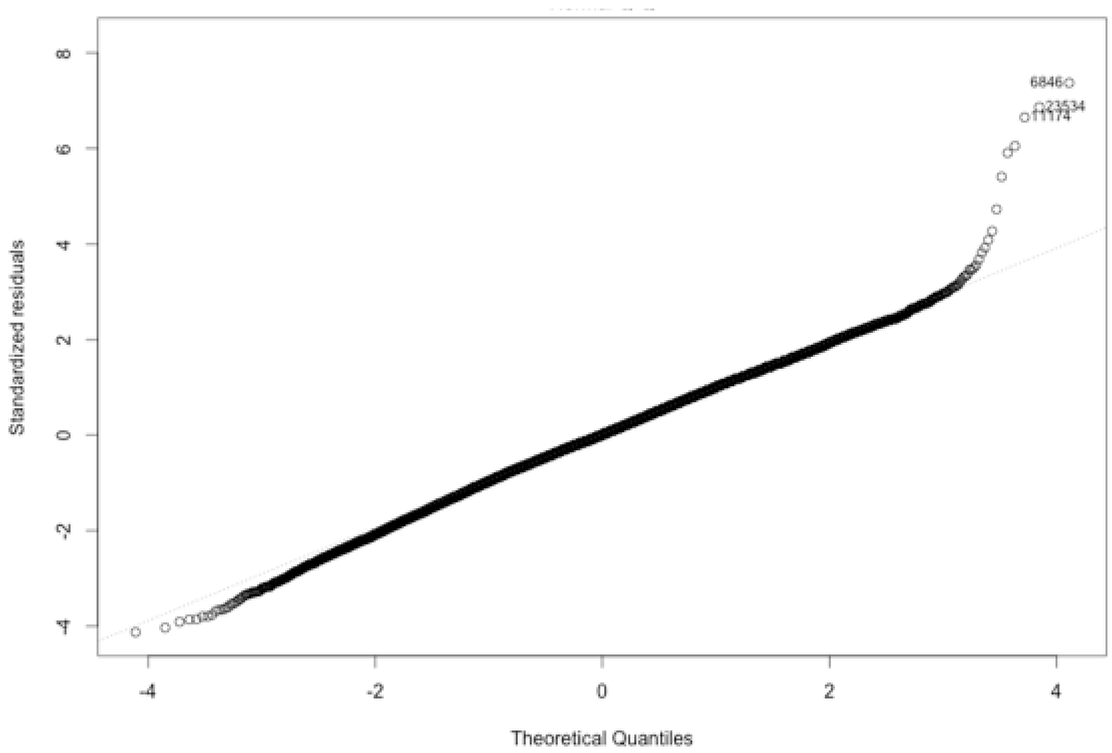
Table 5. Table of Coefficients Estimates (IRS Estimated Population Size)



	Value
Intercept	-6.14
Slope	0.96
Residual Standard Error	0.66
R-squared	0.83
Residual Degrees of Freedom	25378.00

The table of coefficients is for the linear model fit using the logarithmic transformation of the number of leads and the logarithmic transformation of the IRS estimated population size.

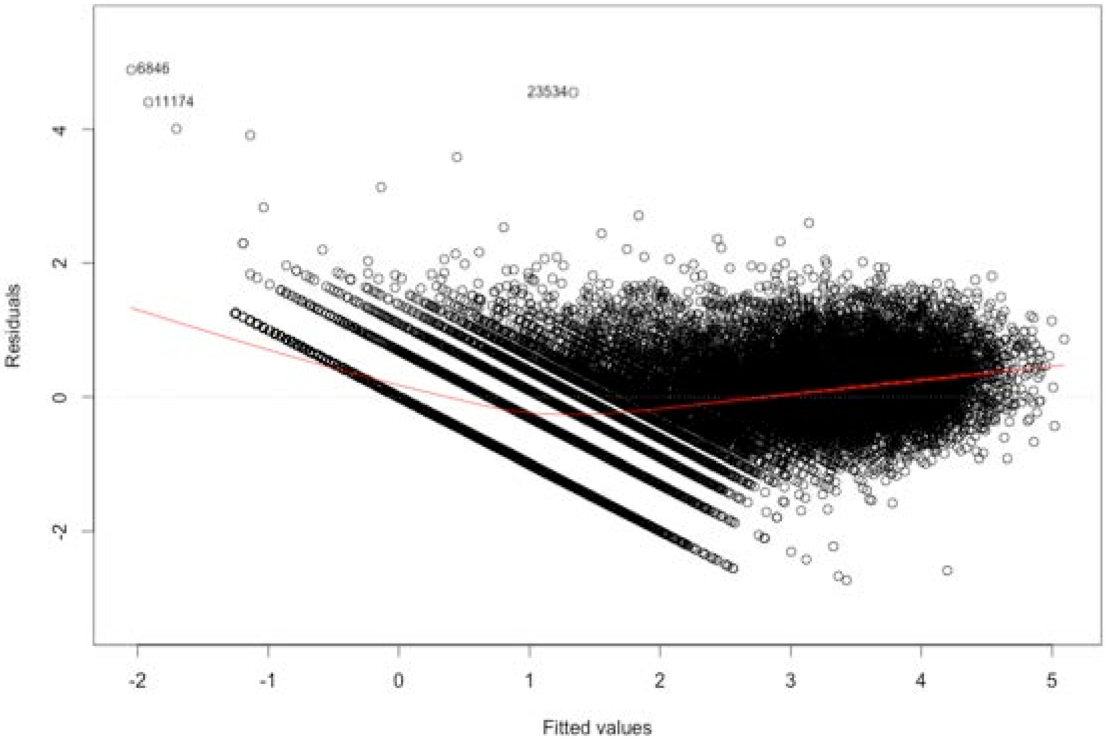
We can see from Table 5 that 82.9% of the variability in the logarithm of the number of leads is explained by a linear relationship with the logarithm of the IRS estimated population size. The plots of the normal Quantile-Quantile (Q-Q) plot of the standardized residuals and the standardized residuals against the fitted values for the model fit in Table 5 are given in Figure 9 and 10 respectively.



The model uses the logarithmic transformation of the number of leads and the logarithmic transformation of IRS estimated population size.

Figure 9. Normal Q-Q Plot of the Standardized Residuals from the Linear Regression Using the Logarithm of IRS Estimated Population Size

Visual inspection of the Figure 9 shows that the residuals are slightly right skewed with eight ZIP codes whose standardized residuals are greater than four, but no significant evidence of non-normality in the residuals. We then review the standardized residuals against the fitted values for the model fit, seen in Figure 10. Figure 10 reveals a non-linear relationship between the logarithm of the number of leads and the logarithm of IRS estimated population size. The red line in Figure 10 is a nonparametric smoother fit to the residuals versus the fitted values.



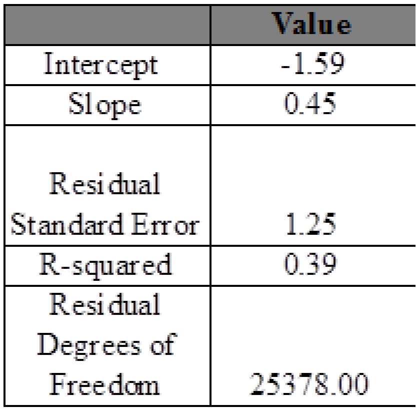
The residuals versus fitted plot from the linear regression of the logarithm of the number of leads against the logarithm of the IRS estimated population size.

Figure 10. Residual versus Fitted Values (IRS Estimated Population Size)

The linear model tends to underestimate the expected logarithm of the number of leads for smaller fitted values and then overestimate as fitted values increase. This finding is not surprising because other characteristics of a region also influence the number of leads generated in a ZIP code. However, the residuals in Figure 10 do not show a dramatic pattern of nonlinearity and the combination of Figures 9 and 10 indicate that the residuals have a nearly normal distribution without an indication of heteroscedasticity. This is a promising sign that the IRS estimated population size could serve as a reasonable proxy for the size of the population generating national leads.

In contrast, the linear model of the logarithm of the number of leads using the logarithm of the Census estimated population size gives an R-squared value of only 39%, as seen in Table 6.

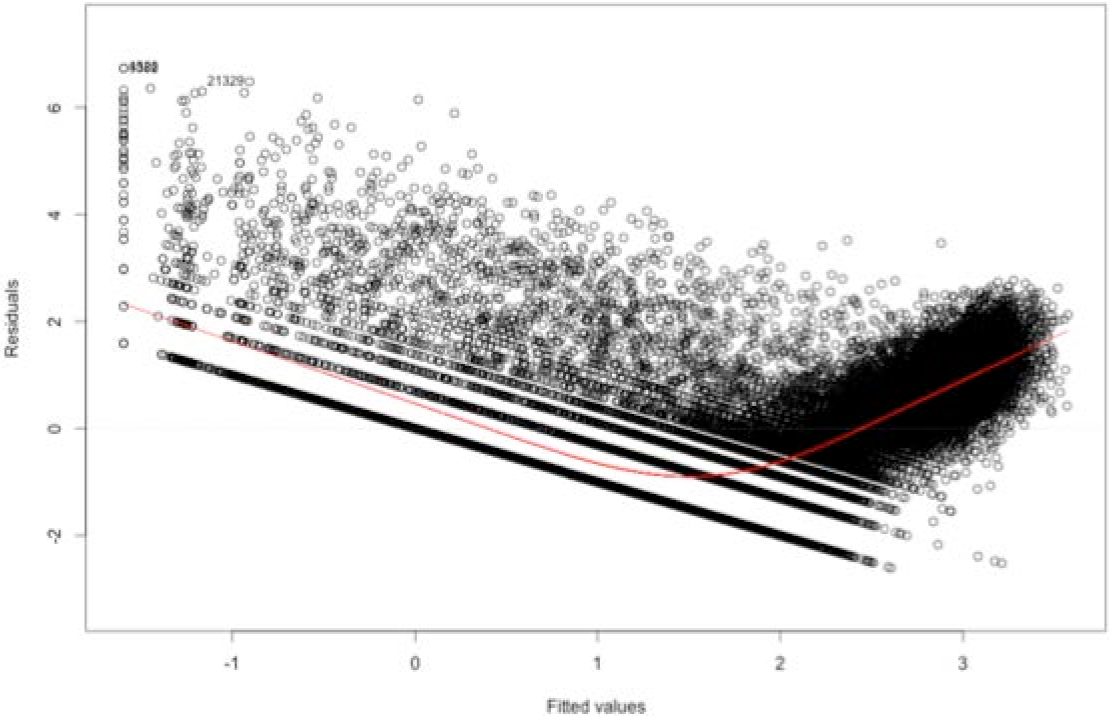
Table 6. Table of Coefficient Estimates (Census Estimated Population Size)



	Value
Intercept	-1.59
Slope	0.45
Residual Standard Error	1.25
R-squared	0.39
Residual Degrees of Freedom	25378.00

The table of coefficients is for the linear model fit using the logarithmic transformation of the number of leads and the logarithmic transformation of the Census estimated population size.

The difference between the R-squared values between the Census and IRS estimated population size leads us to believe that the linear relationship of the logarithm of the number of leads is not explained well by the logarithm of the Census estimated population size. We do a visual inspection of the structure of the model using the residuals versus fitted plot for the Census estimated population size, seen in Figure 11. Here, the red line is the nonparametric smoother and indicates a greater bias than IRS estimated population size.



The residuals versus fitted plot from the linear regression of the logarithm of the number of leads against the logarithm of the Census estimated population size. The red line is a nonparametric smoother fit to the residuals versus the fitted values.

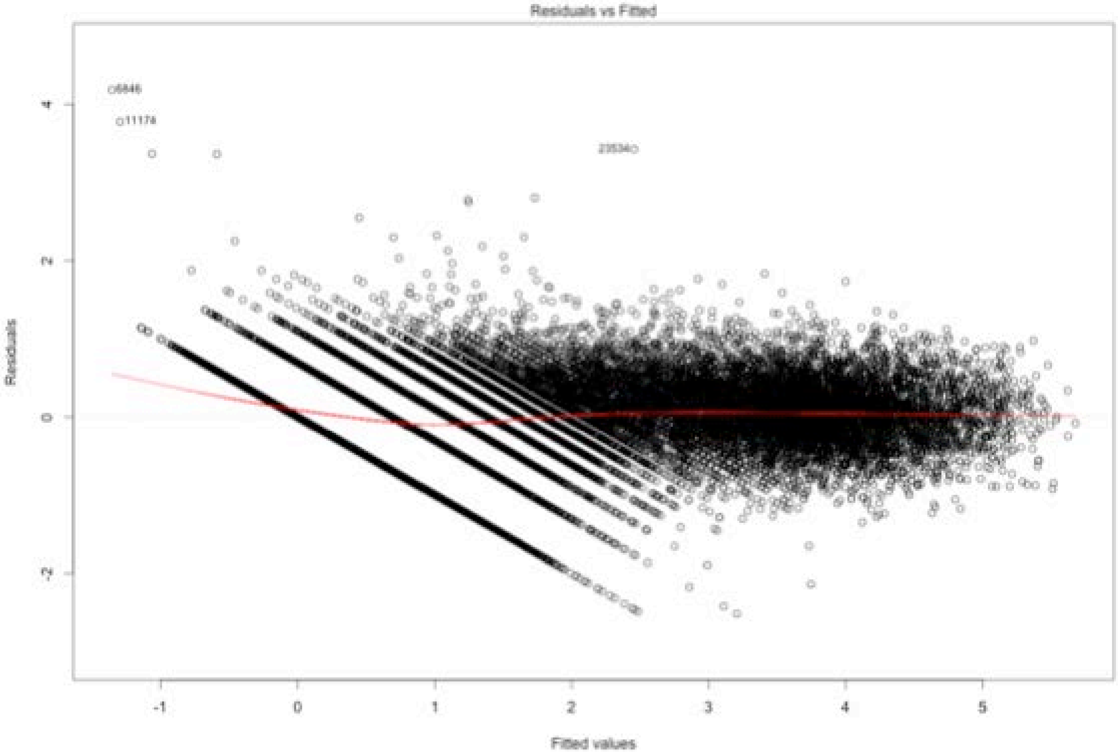
Figure 11. Residuals versus Fitted Values (Census Estimated Population Size)

As the logarithmic transformation of IRS estimated population size explains more of the variability in the number of leads than the Census estimated population size and because the relationship appears to be more linear, we utilize IRS estimated population size in our models rather than the Census estimated population size.

3. ZIP Code Clusters

In order to gain insight into the five *d1* and *d3* Pam generated ZIP code cluster membership factors obtained from Fulton (2016), we use linear regression models with the logarithmic transformation of the number of leads and the logarithm of the IRS estimated population size, adding either *d1* or *d3* Pam generated cluster membership factors.

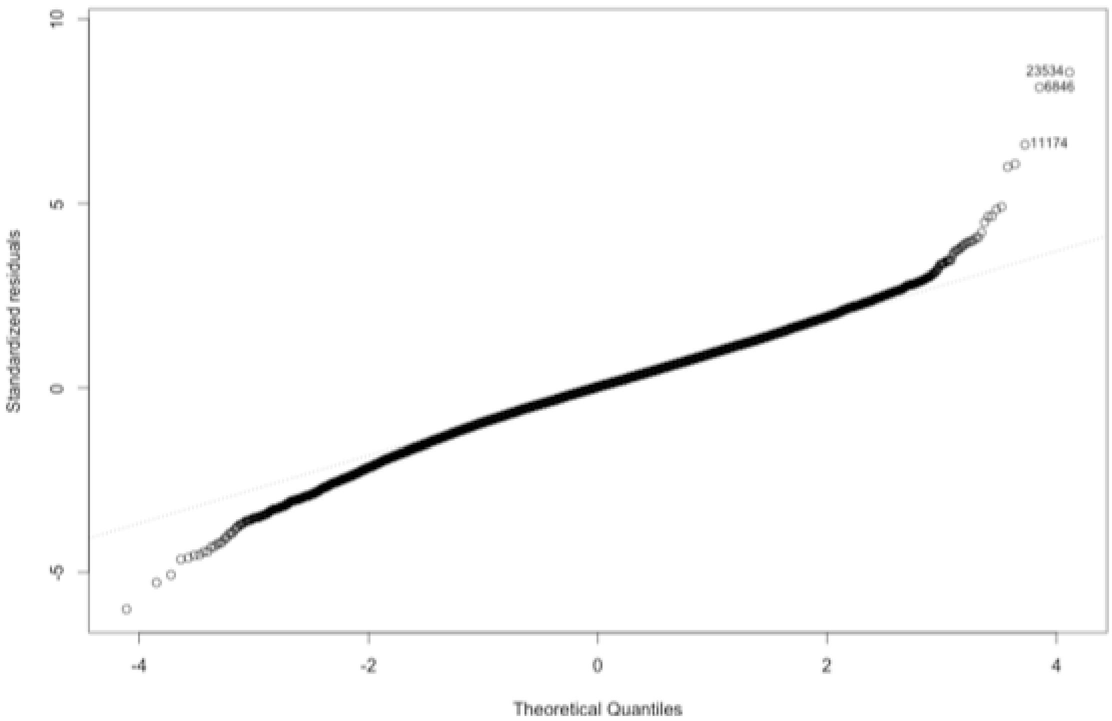
We find that clusters using *d3* Pam results in an R-squared of 89.2% with 0.52 residual standard error on 25320 degrees of freedom, whereas *d1* Pam results in an R-squared of 88.2% with 0.55 residual standard error on 25338 degrees of freedom. We review the residual versus fitted value plots as seen in Figure 12 for *d3* Pam. The red line in Figure 12 is a nonparametric smoother fit to the residuals versus the fitted values.



The residuals versus fitted plot from the linear regression of the logarithm of the number of leads against the logarithm of the IRS estimated population size plus five cluster membership factors.

Figure 12. Residuals versus Fitted Values (*d3* Pam)

Figure 13 gives the normal Q-Q plot of the standardized residuals using *d3* Pam.



The Linear Regression model uses the logarithmic transformation of the number of leads, the logarithmic transformation of IRS estimated population size, and five cluster membership factors.

Figure 13. Normal Q-Q Plot of the Standardized Residuals Using *d3* Pam

Both Figure 12 and 13 suggest a nearly linear relationship and residuals with almost constant variance. The *d1* Pam residual plots are similar. For the rest of our models we use *d3* Pam for cluster membership factors obtained from Fulton (2016), given its slightly higher R-squared value. The most striking feature of Figure 12 are the striations caused by the large number of ZIP codes with no or only a few leads. Keeping this in mind, we begin our model development with a Poisson regression model.

B. MODEL DEVELOPMENT

One of our goals is to keep the model for predicting the number of leads as simple as possible. The analysis of each model focuses on the model’s goodness of fit and what the model indicates as important factors in gaining leads in a ZIP code.

1. Poisson Regression Model

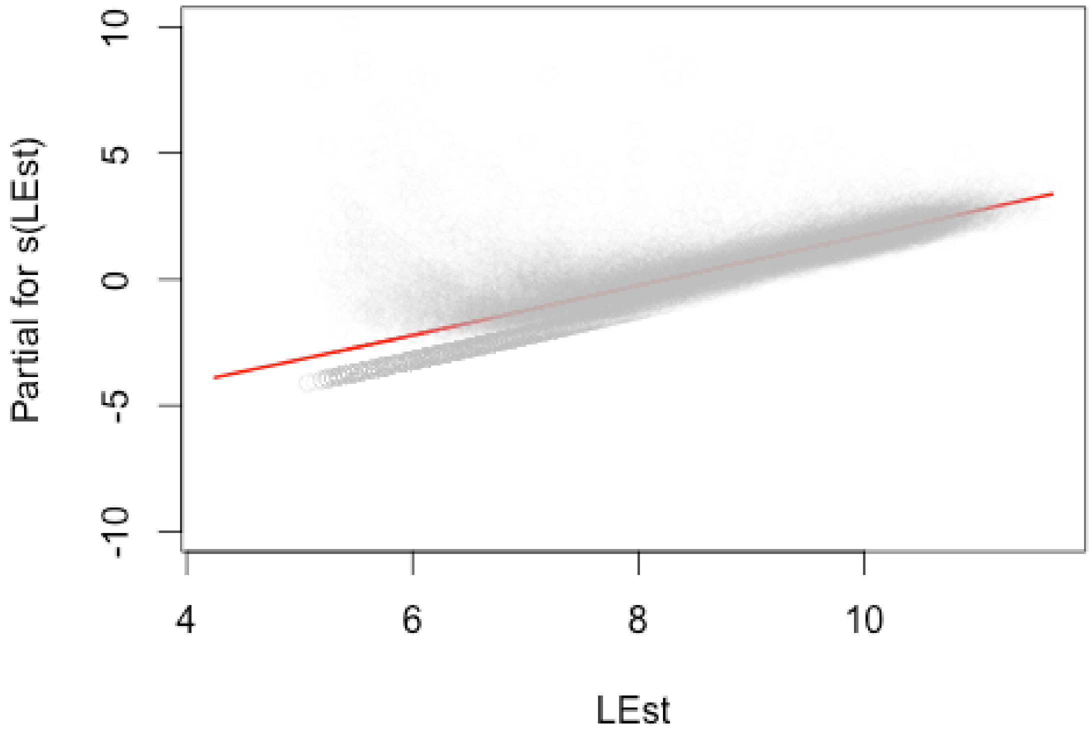
As an initial model we fit a Poisson regression model using the number of leads as the response variable and, as predictors, the five cluster membership factors obtained from Fulton (2016) and the logarithm of IRS estimated population size.

The simple Poisson regression model does not include number of QMA by PRIZM NE segments and results in a simple model with good explanatory power. Using backwards elimination, we remove variables with a p-value greater than 0.05, in this case only one variable, the education cluster membership factor. The model explains 90.1% of the deviance in the number of leads, with a residual deviance of 104199 on 25326 residual degrees of freedom. The coefficient of the logarithm of IRS estimated population size of 0.99 is close to 1, suggesting that an increase in the estimated population size is associated with an expected increase in the predicted the number of leads over a four-year period. This is what we would expect when the estimated population size in acts as a substitute for the number of opportunities to generate leads in a ZIP code.

We include the model output in Appendix A for reference. The table in Appendix A includes standard errors and p-values corresponding to the usual large sample tests of the coefficients, but these standard errors and p-values are biased low since this model does not include interaction terms or other variables nor adjusts for over-dispersion as does the model in the next section.

2. Models with Transformations and Interactions

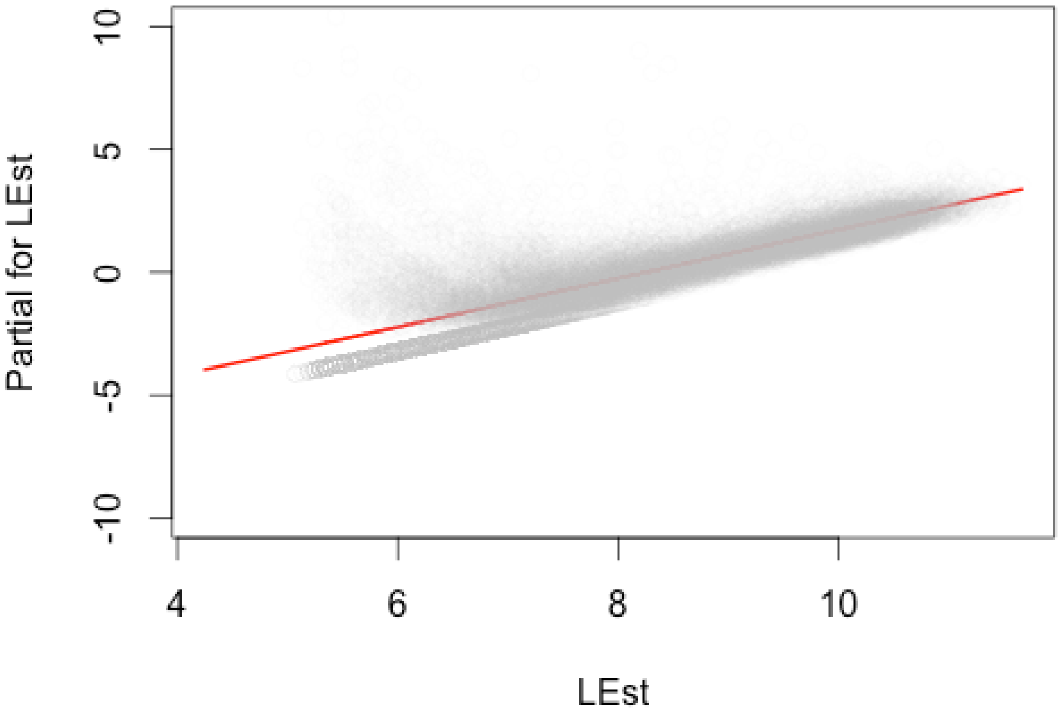
The development of the model with interactions uses our initial Poisson regression model as a starting point. First we examine whether the numeric predictor of IRS estimated population size should be transformed. We fit a generalized additive model with the number of leads as the response variable. The generalized additive model estimates a smooth function of the logarithm of the IRS estimated population size and includes the five cluster membership factors. Figure 14 depicts the partial residual plot for the logarithm of the IRS estimated population size from the generalized additive model fit.



This Poisson generalized additive model utilizes the number of leads as the dependent variable and the smooth function of the logarithmic transformation of the IRS estimated population size (s(LEst))and five cluster membership factors as independent variables.

Figure 14. Partial Residual Plot for the Smooth Function of the Logarithmic Transformation of IRS Estimated Population Size

We compare Figure 14 to Figure 15, the partial residual plot of the Poisson regression model utilizing the logarithmic transformation of IRS estimated population size and the five cluster membership factors as predictors.



This Poisson regression model utilizes the number of leads as the dependent variable and the logarithmic transformation of the IRS estimated population size (LEst) as well as the five cluster membership factors as independent variables.

Figure 15. Partial Residual Plot for the Logarithmic Transformation of IRS Estimated Population Size

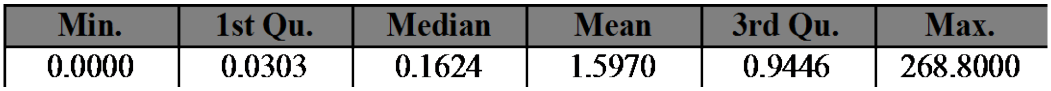
Figure 14 and 15 look almost identical and are a strong indication that a linear partial fit for the logarithm of the IRS estimated population size (as seen in Figure 15) is appropriate. The linear partial fit of the logarithm of the IRS estimated population size, along with the estimated coefficient, suggests that the IRS estimated population size is serving as a "size" variable for the Poisson regression model (Faraway 2006). In this application, size represents the number of opportunities to generate leads in a ZIP code over the four-year period.

Secondly, we investigate the need for extra interaction terms for the five cluster membership factors. We use as predictors: the logarithmic transformation of IRS estimated population size, and five cluster membership factors and their interactions. The model has a residual deviance of 83892 on 24362 to the residual degrees of freedom. It provides an estimate of the dispersion parameter. The dispersion parameter is 1 for a Poisson response variable (Faraway 2006). A dispersion parameter estimate this large indicates over dispersion, which is often caused by unexplained positive dependence among the responses. To adjust standard errors and results of hypothesis tests to accommodate this possible over dispersion, we fit “quasi-Poisson” *glm* models.

The model with interactions explains 92.0% of the deviance, 2% more than the model without interactions. The large sample likelihood ratio test (LRT) of the null hypothesis model without interactions is rejected in favor of the model with interactions (p-value less than 2.2e-16). Backwards elimination removes none of the interaction terms, see Appendix B for results. This is our final model.

Finally, we try adding the number of QMA by PRIZM NE segment, a total of 66 variables, to the final model. Although NRC does not want to use number of QMA per ZIP code to measure of market potential, we wish to see if there is added benefit in predicting the number of leads using the number of QMA by PRIZM NE segment. The resulting pseudo R-squared is slightly higher at 92.9%, but the large sample LRT p-value rejects the final model in favor of the model with the 66 extra variables. This may be an example of statistical significance but not practical significance. The correlation between the fitted values of the two models is 0.985. Summary statistics for the absolute difference between the two models’ predicted number of leads is given in Table 7.

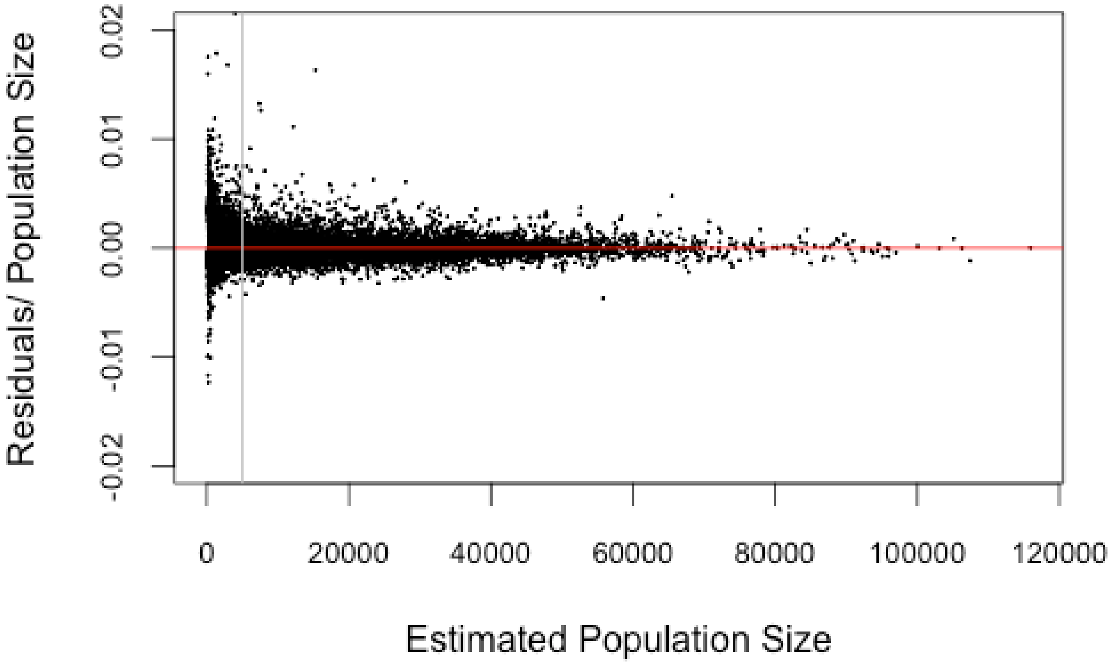
Table 7. Summary Statistics for the Absolute Difference in Predicted Number of Leads



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0303	0.1624	1.5970	0.9446	268.8000

Seventy five percent of the absolute difference in the number of leads is less than one lead with a 90th percentile of 3.8 leads. Only 50 ZIP codes have an absolute difference greater than 50 leads and all of these have a population size greater than 15,000. To keep the model manageable, we use the final model excluding the number of QMA for PRIZM NE segments (66 variables).

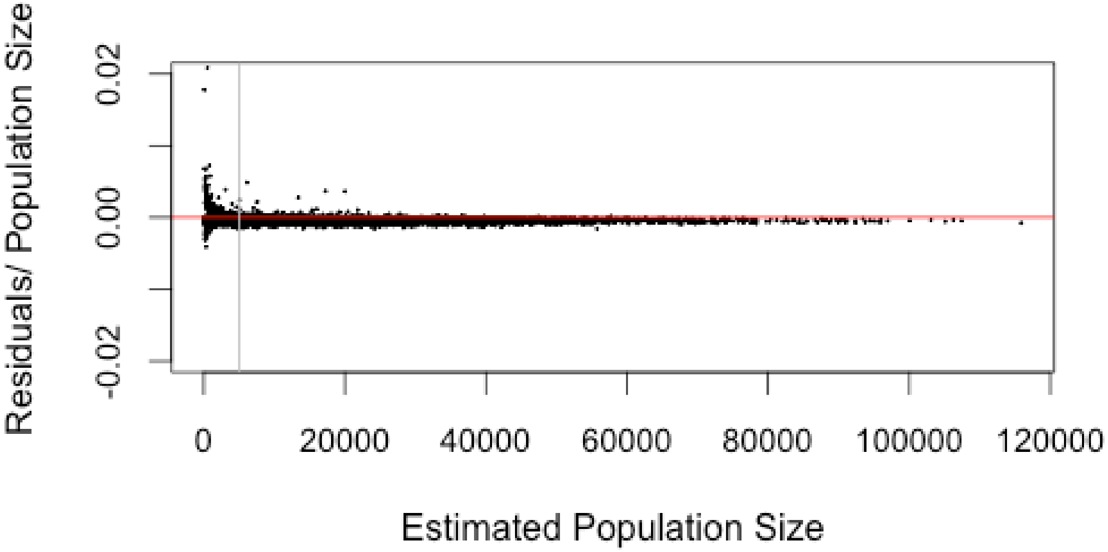
Observations from the training set and the test set are evaluated using the final model with interactions. To account for the population size, in Figures 16 and 17, we divide the residuals (the actual minus the predicted number of leads) by the population size. Figure 16 displays these residuals versus the IRS estimated population size on the training set and Figure 17 does the same for the test set. From our visual inspection of the plot of the training set in Figure 16, the model appears to tend to slightly overestimate the number of leads. Our maximum absolute residual per population size for the training set is 8.7% for all ZIP codes and 1.6% for ZIP codes with an IRS estimated population size greater than 5,000.



The plot of the residuals divided by population size versus the logarithmic transformation of IRS estimated population size on utilizing the training set model on the training set.

Figure 16. Training Set Plot of the Residuals versus IRS Estimated Population Size

As we aggregate the number of leads from FY 11 to FY 14 for the training set, to predict the number of leads for the one year of the test set from the Poisson regression model fit to the total number of leads over four years, we divide the predicted number of leads by four. This gives the plot in Figure 17.



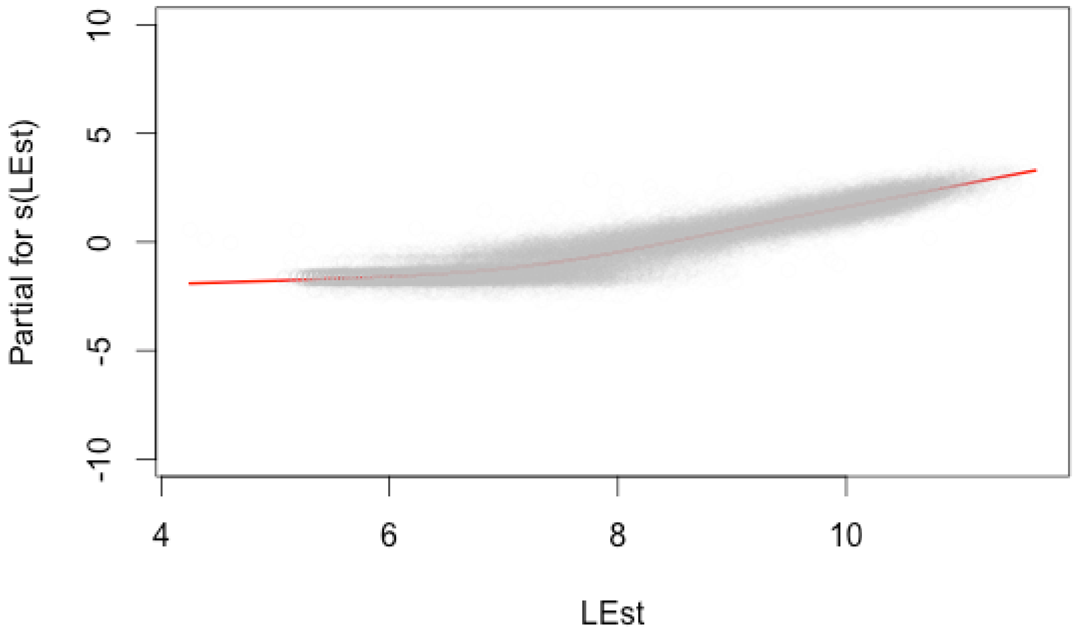
The plot of the residuals divided by population size versus the logarithmic transformation of IRS estimated population size on utilizing the training set model on the test set.

Figure 17. Test Set Plot of the Residuals versus IRS Estimated Population Size

Visually inspecting Figure 17, we can see that the Poisson model with interactions overestimates the number of FY 15 leads as expected. Our maximum absolute residual per population size for the test set is 0.9% for all ZIP codes and 0.05% for ZIP codes with an IRS estimated population size greater than 5,000. Given the bias in the ZIP codes with a larger population size, perhaps leads are easier to generate in ZIP codes with larger population sizes.

3. Multiple Regression Model

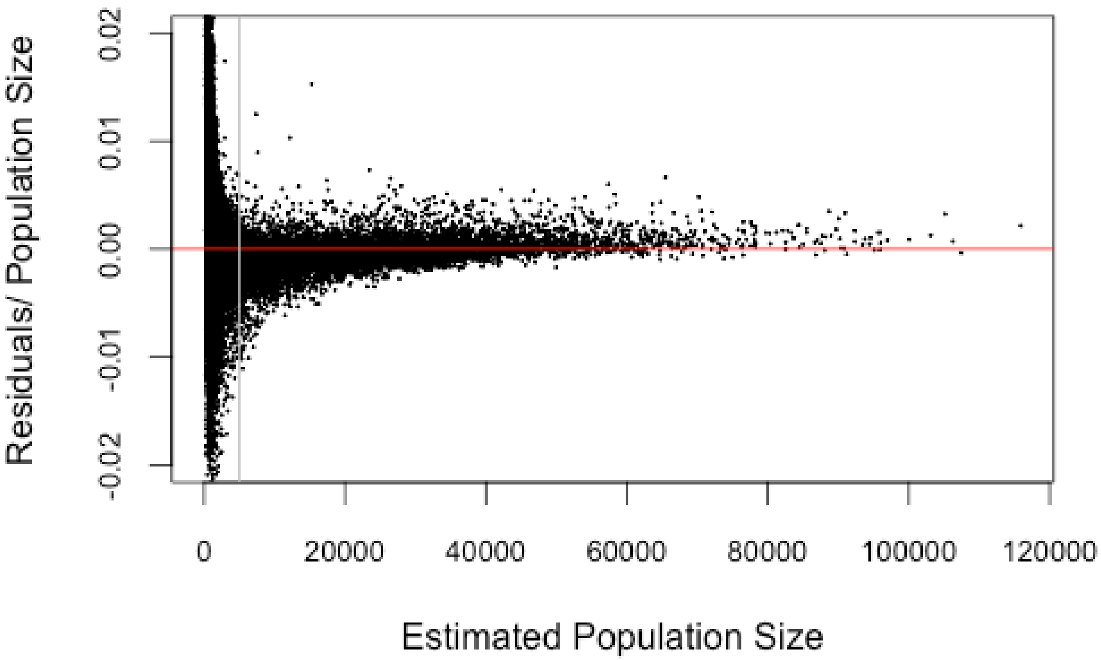
As multiple linear regression models are more familiar and are easier to use we attempt to fit a multiple regression model. The multiple linear model development begins with the logarithmic transformation of the number of leads as the response variables, and as predictors a smooth function of the logarithm of the IRS estimated population size and five cluster membership factors obtained from Fulton (2016) and their interactions. We look at the estimated smooth partial fit for the logarithm of IRS estimated population size, seen in Figure 18.



This generalized additive model utilizes the logarithmic transformation of the number of leads as the response and as predictors: a smooth function of the logarithmic transformation of the IRS estimated population size (seen on the y-axis label as s(LEst)), and the five cluster membership factors.

Figure 18. Partial Residual Plot for the Logarithmic Transformation of IRS Estimated Population Size

The nonlinearity of the partial fit in Figure 18 suggests that a different transformation of IRS estimated population size, for example, the square root transformation often used for Poisson regression, be used for the multiple regression model. Further, the model results in a R-squared value of 63.8%. Figure 19 depicts the plot for the residuals versus the IRS estimated population size of the training set. The red line indicates zero and the gray line on the plot depicts where IRS estimated population size equals 5,000.



The plot of the residuals, in this case the residuals divided by estimated population size, versus the IRS estimated population size on the training set (FY 11 to FY 14) model.

Figure 19. Plot of the Residuals versus IRS Estimated Population Size

The populations smaller than 5,000 show more variability in the residuals than in the Poisson model fit. In addition the model tends to over-estimate the number of leads. The maximum absolute residual per population size for the training set is 1.52% for populations greater than 5,000 and 37.3% for all population sizes. In general, the Poisson regression model fits better than the multiple regression model fit and is a more natural model for count response variables.

V. SUMMARY AND CONCLUSION

A. SUMMARY

Our intent in this research is to develop a statistical model that can predict a geographical area's number of leads, as a measure of market depth, without utilizing the number of past accessions or number of recruiters. In order to accomplish this goal, two questions for analysis are addressed within this research:

- What factors influence the number of national leads?
- Can we predict the number of leads with publicly available data?

This research aggregates leads to the ZIP code level from FY 11 to FY 14 and uses five cluster membership factors obtained from Fulton (2016) to describe respectively the economic, education, military, health, and demographic characteristics of each ZIP code. These five factors represent ZIP code cluster memberships based on a total of 347 variables from publicly available sources. Finally, we developed models to attempt to provide NRC with the information necessary to measure market depth.

The models we develop, Poisson regression models, can predict a geographic area's ability to produce leads. These models only require the logarithmic transformation of IRS estimated population size and the five cluster membership factors, but their predictive ability may be slightly improved with the addition of the 66 variables representing the number of QMA for each PRIZM NE segment.

This research also demonstrates that IRS estimated population size is a better proxy for the number in the population who might produce a lead than Census estimated population size mapped to the ZIP code-level.

B. RECOMMENDATIONS

First, we recommend that NRC implement the Poisson model with interactions as a tool for understanding market depth. This model should not replace models that calculate RAF or best predict accessions but should be used to identify areas with the potential for more leads. The implementation of the Poisson model will allow the NRC to identify ZIP codes with a high probability of greater market depth, as shown by a greater potential for more leads.

Second, we recommend that NRC use national leads as one measure of market depth. We hypothesize that the work of this research might then support future work regarding recruiters using person-to-person video chat, such as Skype or Face Time, rather than working locally.

Finally, we recommend that number of QMA by PRIZM NE segmentation data, if used, be used for its original intention, for marketing at the ZIP code level, rather than for predictive analysis.

C. FUTURE WORK

Future analysis in this area should focus on further refinement of IRS estimated population size, specifically utilizing estimated youth population size. To follow on with utilizing estimated youth population size, future analysis in this area should focus on utilizing the number of leads across the DOD in order to gain further insights into geographic areas that may have service-specific influences.

There appear to be real changes in the number of leads over time that should be investigated further. The data in support of this research does not utilize Armed Services Vocational Aptitude Battery scores for potential accessions, a Navy Awareness Index utilized by Pinelis et al. (2011), or conversion rates to accessions by ZIP code. By addressing these areas, future analysts might be better able to model market depth as a function of the number of leads.

APPENDIX A. POISSON REGRESSION MODEL

This appendix contains the full Poisson regression model output produced by the *stats* package in R. This output provides any additional statistics not included in the body of this research. The standard errors and p-values seen here are biased low since this model does not include interaction terms nor adjusts for over dispersion, as does the model in the next Appendix. Note that ‘LEst’ is the logarithm of the IRS estimated population sizes and that e.g., “factor(econ)2” represents the indication function for ZIP code cluster membership in cluster 2 based on clustering ZIP codes with the economic variables. The coefficients for the cluster 1 indication functions are zero.

```
Call:
glm(formula = freq ~ (LEst) + (factor(health) + factor(econ) + 
    factor(ed) + factor(mil) + factor(demo)), family = "poisson", 
    data = n)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-15.659   -1.261   -0.455    0.697   42.165  

Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)    
(Intercept)      -6.476339   0.029132 -222.312  < 2e-16 ***
LEst              0.990073   0.002702  366.405  < 2e-16 ***
factor(health)2   0.000893   0.007864    0.114 0.909589    
factor(health)3   0.276110   0.007627   36.203  < 2e-16 ***
factor(health)4  -0.223644   0.009230  -24.229  < 2e-16 ***
factor(health)5  -0.159121   0.007063  -22.530  < 2e-16 ***
factor(health)6  -0.144878   0.009139  -15.853  < 2e-16 ***
factor(health)7  -0.234400   0.014601  -16.054  < 2e-16 ***
factor(health)8   0.169310   0.007819   21.654  < 2e-16 ***
factor(health)9   0.081046   0.006675   12.142  < 2e-16 ***
factor(health)10  0.083231   0.008285   10.046  < 2e-16 ***
factor(health)11 -0.062168   0.010113   -6.148 7.87e-10 ***
factor(health)12  0.024980   0.007227    3.456 0.000547 ***
factor(econ)2     0.182106   0.012413   14.671  < 2e-16 ***
factor(econ)3     0.012955   0.026638    0.486 0.626725    
factor(econ)4     0.057699   0.013097    4.405 1.06e-05 ***
factor(econ)5     0.570051   0.012213   46.677  < 2e-16 ***
factor(econ)6     0.275238   0.012451   22.105  < 2e-16 ***
factor(econ)7     0.409844   0.012485   32.828  < 2e-16 ***
factor(econ)8    -0.082259   0.040651   -2.024 0.043017 *  
factor(econ)9    -0.146557   0.026874   -5.453 4.94e-08 ***
factor(econ)10    0.123087   0.012977    9.485  < 2e-16 ***
factor(econ)11   -0.083769   0.015686   -5.340 9.28e-08 ***
factor(econ)12    0.326838   0.012352   26.460  < 2e-16 ***
factor(econ)13    0.038117   0.031851    1.197 0.231418    
factor(econ)14   -0.280884   0.045167   -6.219 5.01e-10 ***
factor(econ)15   -0.004671   0.012995   -0.359 0.719279    
factor(econ)16    0.304976   0.020242   15.066  < 2e-16 ***
factor(econ)17    0.291223   0.013095   22.240  < 2e-16 ***
factor(ed)2      -0.009720   0.003872   -2.510 0.012064 *  
factor(mil)2     -0.060634   0.010292   -5.892 3.83e-09 ***
factor(mil)3     -0.155616   0.010169  -15.303  < 2e-16 ***
factor(mil)4      0.191457   0.009883   19.373  < 2e-16 ***
factor(mil)5     -0.066639   0.010628   -6.270 3.61e-10 ***
factor(mil)6      0.211230   0.009999   21.125  < 2e-16 ***
factor(mil)7      0.034279   0.009992    3.431 0.000602 ***
factor(mil)8      0.127604   0.013087    9.751  < 2e-16 ***
factor(mil)9      0.101582   0.010584    9.598  < 2e-16 ***
factor(mil)10     0.493260   0.009133   54.010  < 2e-16 ***
factor(mil)11     0.251154   0.010718   23.432  < 2e-16 ***
factor(mil)12     0.196926   0.012122   16.246  < 2e-16 ***
factor(mil)13     0.153044   0.010850   14.105  < 2e-16 ***
factor(mil)14    -0.020219   0.016088   -1.257 0.208840    
factor(mil)15    -0.077724   0.018906   -4.111 3.94e-05 ***
factor(mil)16     0.159370   0.009347   17.051  < 2e-16 ***
factor(demo)2    -0.215951   0.007472  -28.900  < 2e-16 ***
factor(demo)3    -0.114177   0.008593  -13.287  < 2e-16 ***
factor(demo)4    -0.292031   0.009485  -30.790  < 2e-16 ***
factor(demo)5     0.195412   0.004886   39.997  < 2e-16 ***
factor(demo)6    -0.122393   0.005052  -24.227  < 2e-16 ***
factor(demo)7    -0.214085   0.008090  -26.463  < 2e-16 ***
factor(demo)8    -0.273975   0.020891  -13.114  < 2e-16 ***
factor(demo)9    -0.160375   0.007045  -22.764  < 2e-16 ***
factor(demo)10    0.125358   0.006939   18.066  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1053670  on 25379  degrees of freedom
Residual deviance:  104199  on 25326  degrees of freedom
AIC: 189346

Number of Fisher Scoring iterations: 5
```

APPENDIX B. MODEL WITH INTERACTIONS

This appendix contains the output for the final model with interactions produced by the *stats* package in R. This output provides any additional statistics not included in the body of this research. See Appendix A for an explanation of notation. Here “:” indicates the interaction between the variables on either side of the colon. The table given here shows the results of the large sample LRT for each interaction term in the presence of the rest.

```
Model:
glm(freq ~ LEst + (factor(health) + factor(econ) + factor(mil) +
    factor(demo) + factor(ed))^2, family=”quasipoisson”)
                             Df Deviance    F value    Pr(>F)
<none>                             83892
LEst                          1   207954 36027.2753 < 2.2e-16 ***
factor(health):factor(econ) 175    85047     1.9161 4.720e-12 ***
factor(health):factor(mil)  147    86644     5.4369 < 2.2e-16 ***
factor(health):factor(demo)  84    85957     7.1391 < 2.2e-16 ***
factor(health):factor(ed)    11    84386    13.0537 < 2.2e-16 ***
factor(econ):factor(mil)    236    85504     1.9831 < 2.2e-16 ***
factor(econ):factor(demo)   140    85329     2.9807 < 2.2e-16 ***
factor(econ):factor(ed)      16    84061     3.0762 3.090e-05 ***
factor(mil):factor(demo)    129    85516     3.6551 < 2.2e-16 ***
factor(mil):factor(ed)       15    84202     6.0003 1.052e-12 ***
factor(demo):factor(ed)       9    84062     5.4953 1.388e-07 ***
---
Signif.  codes:   0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘  ’  1
(Dispersion parameter for quasipoisson family taken to be 4.037366)

    Null deviance: 1053670  on 25379  degrees of freedom
Residual deviance:   83892  on 24362  degrees of freedom
```

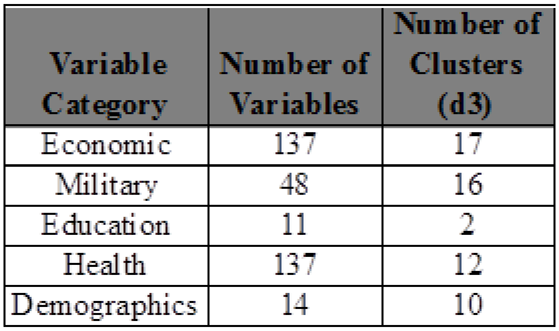
THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. MAPPING CLUSTERS

This appendix contains the map output for each of the five cluster membership factors obtained from Fulton (2016), which were clustered using publicly available variables. This appendix provides any additional map information for clusters not included in the body of this research.

As discussed in the body of this thesis, Fulton (2016) uses several different clustering algorithms and dissimilarity calculations. This appendix maps each variable category for *d3* Pam, the clusters primarily used in this research. Table 8 shows each variable category, the number of variables, and the number of associated clusters.

Table 8. *d3* Pam Cluster Assignments for each Variable



Variable Category	Number of Variables	Number of Clusters (d3)
Economic	137	17
Military	48	16
Education	11	2
Health	137	12
Demographics	14	10

Each category uses a different number of clusters, determined by the smallest average normalized distance from the medoid of the cluster by Fulton (2016).

The economic factor requires a small amount of data preparation, as one of the ZIP codes does not have a cluster assignment. This data preparation is described in detail in Chapter III. We can see the map of the economic factor in Figure 20. Clusters 1 through 17 are depicted with red, blue, green, orange, purple, gray, dark green, wheat, royal blue, pink, dark cyan, yellow, cyan, violet, brown, black, and dark red respectively.

Looking at the map, the color yellow for cluster 8 stands out. There are 4,768 ZIP codes classified as cluster 8 for the economic factor, the majority of which are found in California, with 443 ZIP codes. This cluster is characterized by a very low percentage of the population with an average income under 25,000 (0.04%), a high percentage of the population with an average income between 75,000 and 100,000 (15.8%), a high percentage of the population with an average income between 100,000 and 200,000 (10.3%), and a high percentage of the population with an average income over 200,000 (9.2%).

Next, the color pink stands out, which corresponds to cluster 10. There are 4,005 ZIP codes classified as cluster 10 and are found in Pennsylvania, 386. This cluster can be characterized as having a high percentage of the population with an average income under 25,000, in comparison to the other clusters (4%).

The dark red color also stands out, which corresponds to cluster 17. There are 2,909 ZIP codes classified as cluster 17, the majority of which are found in Texas, 267. This cluster can be characterized as having a high percentage of the population with an average income between 25,000 and 50,000 (42.6%).

Finally, the purple color stands out which corresponds to cluster 5, which makes up 1,138 of the ZIP codes, the majority of which are found in Texas, 250. This cluster can be characterized as having the lowest percentages of the population with income from 75,000 to 200,000 in comparison to the other clusters (19%), and a very high average population making 25,000 to 50,000 (45.6%).

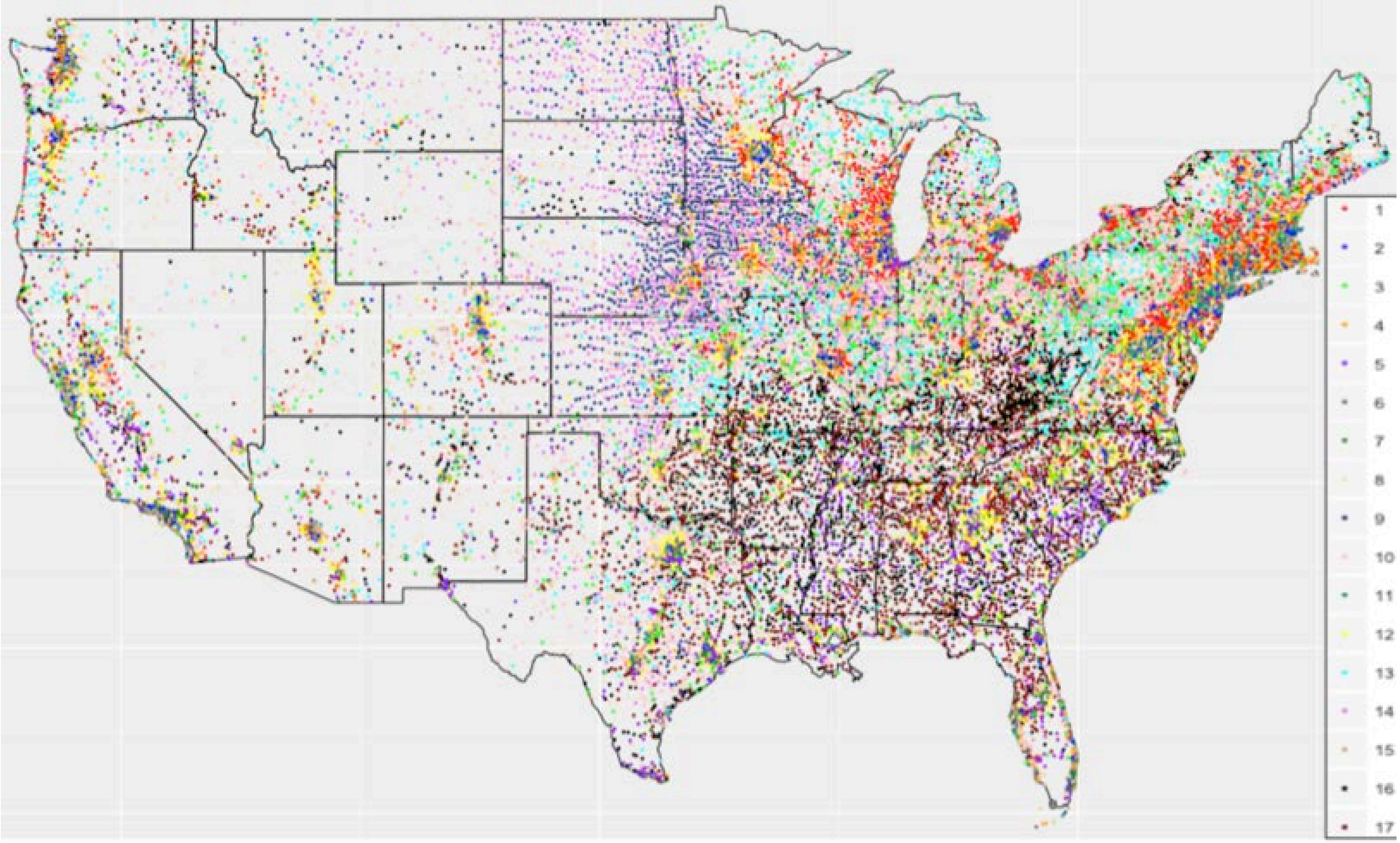


Figure 20. Map Showing Cluster Assignments of the Economic Factor Using *d3* Pam

Next we look at the military factor. The military factor utilizes 16 clusters to achieve the smallest average distance, as seen in Figure 21. Clusters 1 through 16 are depicted with red, blue, green, orange, purple, gray, dark green, wheat, royal blue, pink, dark cyan, yellow, cyan, violet, brown, and black respectively. The colors of purple, wheat, violet, and black immediately stand out when looking at the map of military variables, which corresponds to cluster 5, 14, 8, and 16 respectively.

First, the color purple stands out, which corresponds to cluster 5. There are 3,358 ZIP codes classified as cluster 5. 1,245 ZIP codes of cluster 5 are closest to a depot. 370 ZIP codes in cluster 5 are in Pennsylvania. Cluster 5 can be characterized as having the smallest average population of the nearest military base, approximately 547 people.

The color wheat stands out, which corresponds to cluster 14. 2,778 ZIP codes make up cluster 14. 886 of these ZIP codes are closest to a depot; 828 of these ZIP codes are closest to an Army National Guard site. This cluster is found primarily in Iowa with 219 ZIP codes.

Next, the color violet stands out, which corresponds to cluster 8. There are 2,742 ZIP codes classified as cluster 8. 502 ZIP codes of cluster 8 are closest to a depot; 454 ZIP codes of cluster 8 are closest to an Army National Guard site. The two highest states with this cluster are Pennsylvania (with 432 ZIP codes) and Florida (with 397 ZIP codes). This cluster can be characterized with a high average population of the nearest base, for an average of 990 people.

Finally, cluster 16, in black, has 2,206 ZIP codes. 589 ZIP codes of cluster 16 are closest to an Army National Guard site followed closely by 528 ZIP codes near a depot. This cluster is found primarily in ZIP codes in Texas (with 577 ZIP codes) and California (with 356 ZIP codes). This cluster can be characterized with a short average distance to the nearest military base, approximately 24 miles, with a very high population at that nearest military base, approximately 714 people.

Interestingly, cluster 11, with 851 ZIP codes, has the highest average veteran population of 1,458 people, followed by cluster 10 (1,388 people), and cluster 16 (1,010 people).

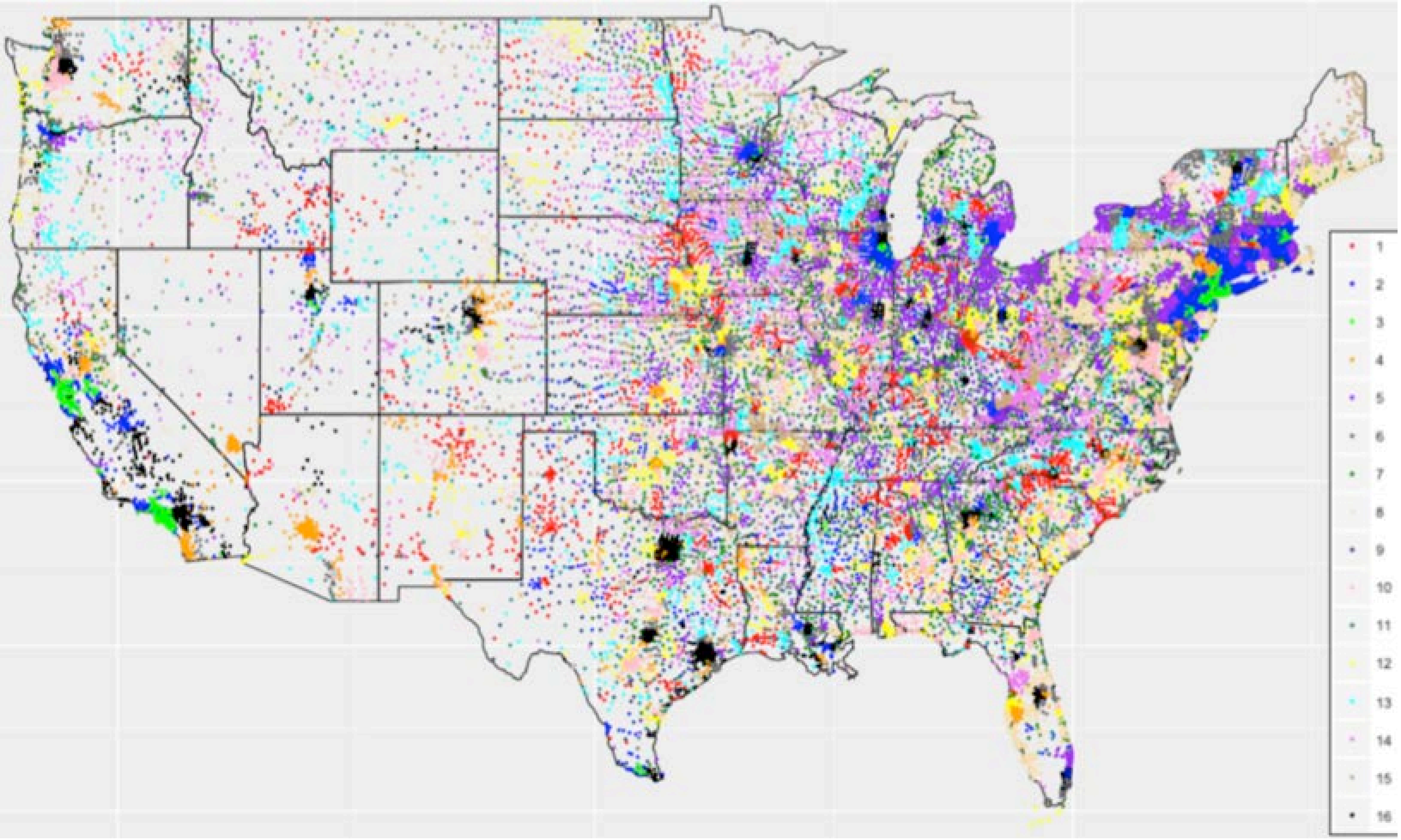


Figure 21. Map Showing Cluster Assignments of the Military Factor Using *d3* Pam

The education factor only utilizes two clusters, as seen in Figure 22. Clusters 1 and 2 are depicted with red and blue respectively. Cluster 1 accounts for 13,104 ZIP codes. Interestingly, the average number of educational institutions smaller than 1,000 people within 10 miles of the ZIP code is 10.8 and the average number of educational institutions smaller than 1,000 people within 50 miles is 83.7 for cluster 1. In every category of educational variables, cluster 2 has a smaller average number of educational institutions than cluster 1, leading us to believe that cluster 2 corresponds with rural areas where secondary education is not as readily available.

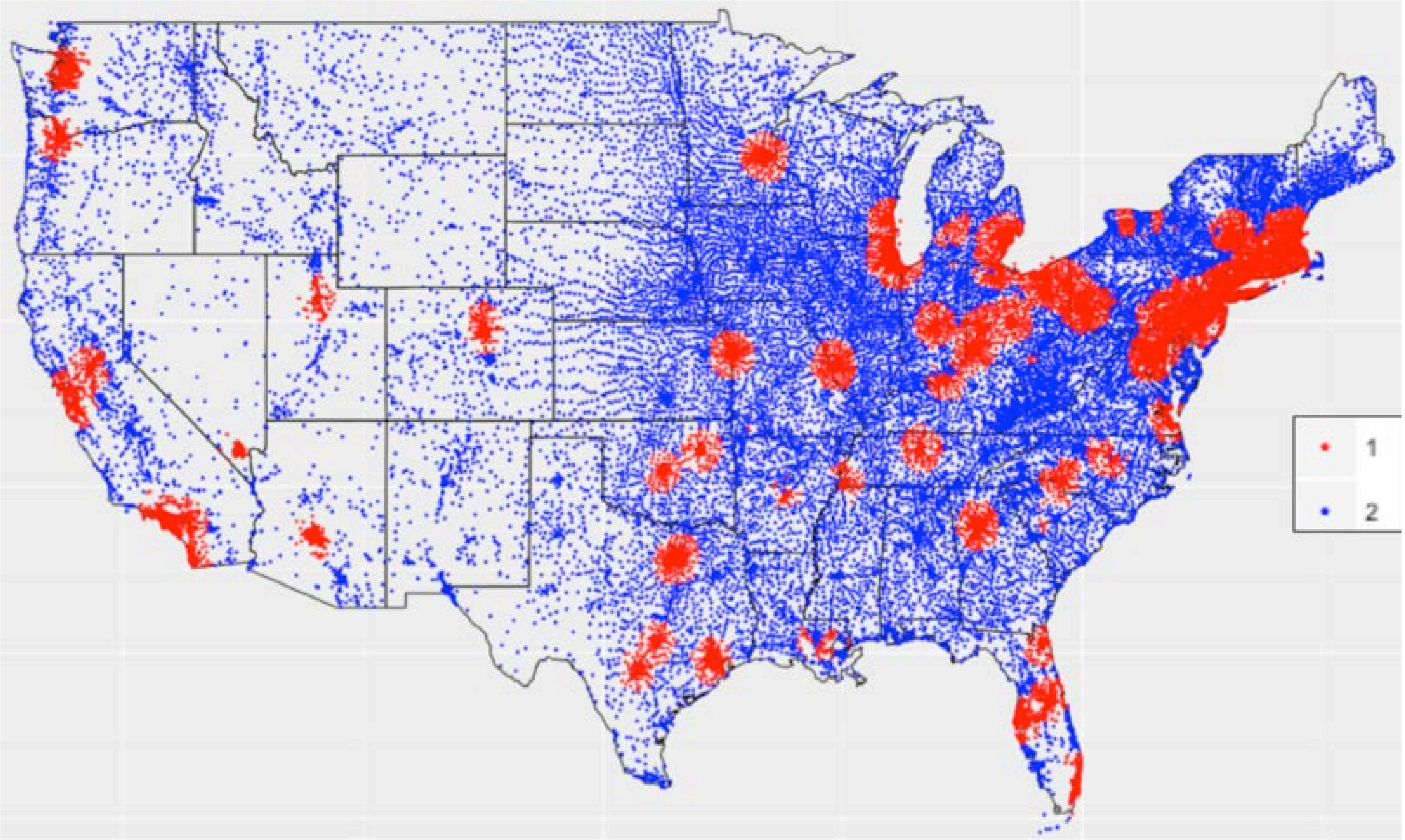


Figure 22. Map Showing Cluster Assignments of the Education Factor Using *d3* Pam.

The health factor uses 12 clusters, as seen in Figure 23. Clusters 1 through 12 are depicted with red, blue, green, orange, purple, gray, dark green, wheat, royal blue, pink, dark cyan, and yellow respectively.

Cluster 4 corresponds to the orange color on the health map and easily stands out. Cluster 4 primarily corresponds to Minnesota, with 485 ZIP codes falling into the cluster, and Wisconsin (with 456 ZIP codes).

Cluster 2 corresponds to the blue color on the health map. New York, with 595 ZIP codes, Pennsylvania (with 510 ZIP codes), California (with 500 ZIP codes) and Massachusetts (with 494 ZIP codes) primarily make up cluster 2. Cluster 2 can be characterized as having the highest average rate of all clusters of recent drug users (within the last month) of 6.46, the lowest average rate of all clusters of obesity of 19.7, and the lowest average rate of all clusters of uninsured of 9.6.

Cluster 11 corresponds to the dark cyan color sprinkled throughout the Mid-Atlantic States. Cluster 11 primarily consists of Texas (with 434 ZIP codes), Kentucky (with 391 ZIP codes), and Oklahoma (with 390 ZIP codes). Cluster 11 can be characterized as having the highest average rate of all clusters of major depression of 6.2 and the highest average rate of all clusters for disabled Medicare beneficiaries of 3.7.

Cluster 12 stands out visually in bright yellow, but only 1413 ZIP codes are in this cluster. This cluster primarily consists of ZIP codes in Texas (with 814 ZIP codes), California (with 340 ZIP codes), and Nevada (with 98 ZIP codes). This cluster can be characterized as having the lowest average rate of all clusters of elderly Medicare beneficiaries of 8.84.

Cluster 10 corresponds to pink, seen on the map throughout the southern part of the United States. This cluster primarily consists of ZIP codes in Alabama (with 359 ZIP codes), Louisiana (with 291 ZIP codes), Georgia (with 275 ZIP codes), North Carolina (with 252 ZIP codes), Mississippi (with 222 ZIP codes), Texas (with 211 ZIP codes), Virginia (with 207 ZIP codes), and West Virginia (with 203 ZIP codes). Interestingly, cluster 10 has the lowest average rate of all clusters for life expectancy (74.1 years), the highest average rate of all clusters for causes of death (1,025), the highest average rate of all clusters for obesity (27.7), and the highest average rate of all clusters for diabetes (9.47).

Finally, cluster 1 stands out visually in tight clusters on the health map. Cluster 1 consists primarily in Florida (with 526 ZIP codes), Arizona (with 343 ZIP codes), and California (with 211 ZIP codes). Cluster 1 does not have any outstanding characteristics, but we hypothesize that cluster 1 demographically corresponds to an elderly population.

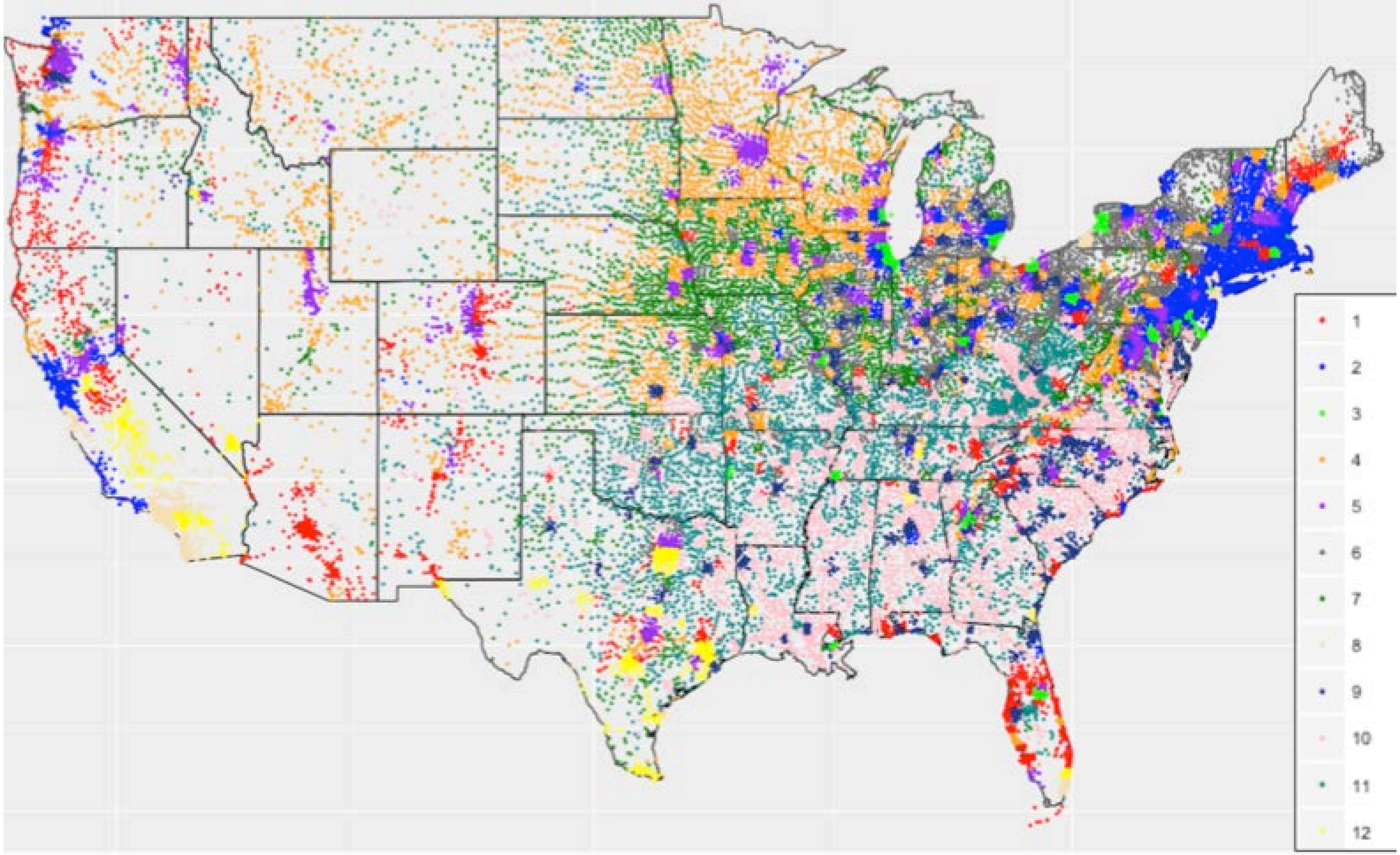


Figure 23. Map Showing Cluster Assignments of Health Variables Using *d3* Pam.

The demographic factor has 10 clusters, as seen in Figure 24. The map of the demographic factor is described in detail in Chapter III. Cluster 1 is red; clusters 2 through 10 are depicted with blue, green, orange, purple, gray, dark green, wheat, royal blue, and pink respectively.

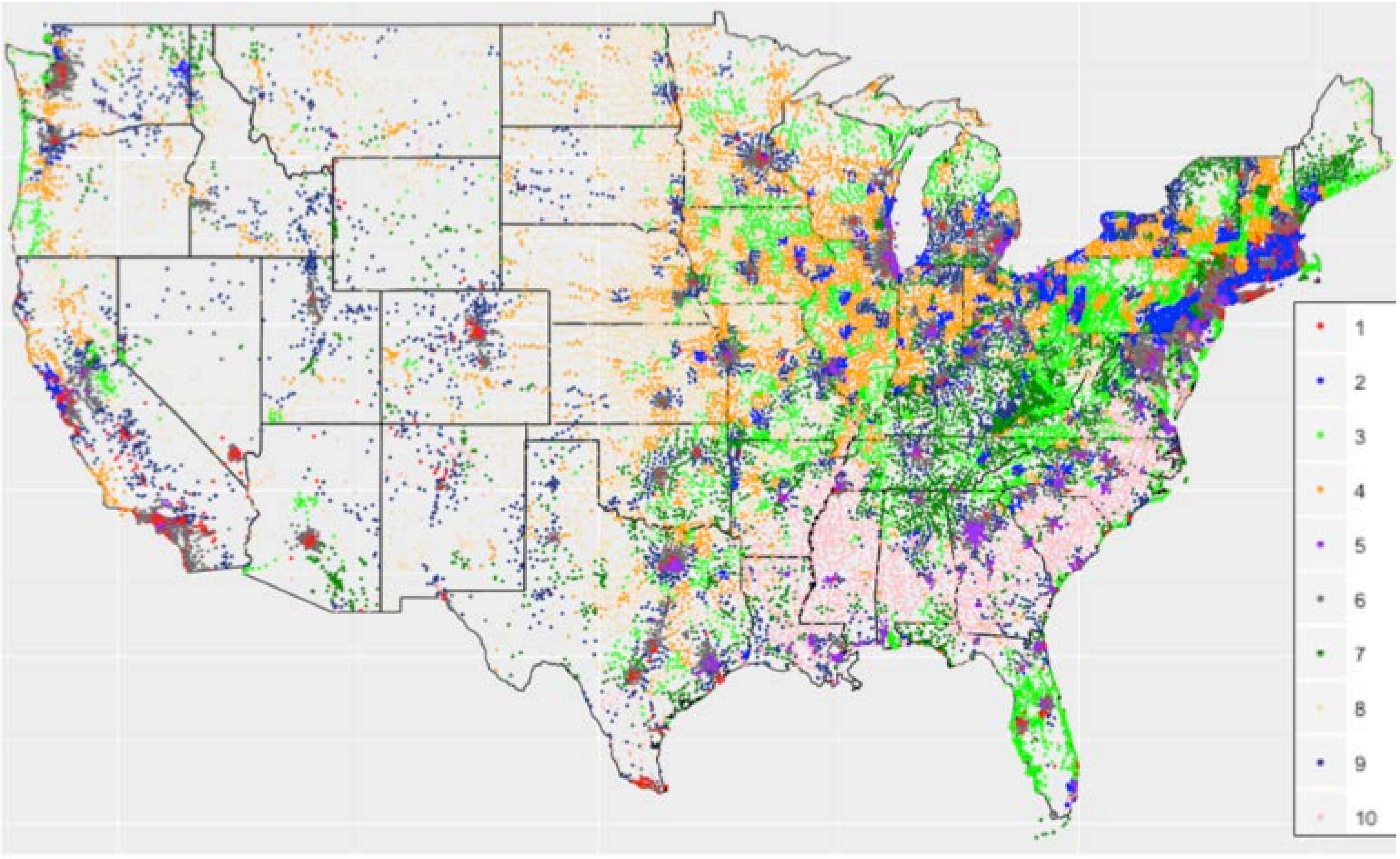


Figure 24. Map Showing Cluster Assignments of the Demographic Factor Using *d3* Pam.

LIST OF REFERENCES

Ammons-Moreno D (2016) E-mail personal communication April 12, 2016.

Buttrey S, Whitaker L (2015) treeClust: an R package for tree-based clustering dissimilarities. *The R Journal* 7/2(December). Retrieved April 8, 2016, https://journal.r-project.org/archive/2015-2/buttrey-whitaker.pdf

Darrow G (2016) Assessing the value of the current propensity metric to U.S. Army Recruiting Command. (Unpublished master's thesis). Personal communication May 31, 2016.

ESRI (2016) ArcGIS Online. Redlands, CA: Environmental Systems Research Institute. Accessed June 2016. Retrieved from http://www.arcgis.com/features/index.html

Faraway J (2006) Extending the linear model with R. Taylor and Francis Group Boca Raton, FL.

Faraway J (2015) Linear models with R. Taylor and Francis Group Boca Raton, FL.

Fulton B (2016) Determining Market Categorization Of United States Zip Codes For Purposes Of Army Recruiting. (Unpublished master's thesis). Personal communication May 29, 2016.

Gibson J, Hermida R, Luchman J, Griepentrog B, Marsh S (2011) ZIP code valuation study technical report, Joint Advertising, Market Research & Studies, Defense Human Resources Activity, Arlington, Virginia.

Hastie T (2015) R package: generalized additive models. London: Chapter 7 of Statistical Models in S.

Housing and Urban Development (2016) Fourth quarter 2015 county to ZIP code crosswalk. Retrieved from https://www.huduser.gov/portal/datasets/usps_crosswalk.html

Internal Revenue Service (IRS) (2016) Internal Revenue Service SOI data. Retrieved May 1, 2016, https://www.irs.gov/uac/SOI-Tax-Stats-Individual-Income-Tax-Statistics-2013-ZIP-Code-Data-(SOI)

Internal Revenue Service (IRS) (2015) Internal Revenue Service SOI tax stats - Individual income tax returns publication 1304: Section 4: Explanation of terms. Retrieved from https://www.irs.gov/pub/irs-soi/13insec4.pdf

Intrater B (2015) Understanding the impact of socio-economic factors on Navy accessions. (Master's thesis) Retrieved from Calhoun http://calhoun.nps.edu/bitstream/handle/10945/47279/15Sep_Intrater_Bradley.pdf

Jordan M (2014) Recruits' ineligibility tests the military – More than two-thirds of American youth wouldn't qualify for service, Pentagon says. *Wall Street Journal* (June 27), Retrieved from http://www.wsj.com/articles/recruits-ineligibility-tests-the-military-1403909945

Larter D and Faram M (2016) Navy personnel chief: 6,300 sailor cuts won't come from force-outs. *Navy Times* (March 3), Retrieved from http://www.navytimes.com/story/military/2016/03/03/cnp-bill-moran-force-out-boards-no-reason-end-strength-cuts/81165632/

Maechler M, Rousseeuw P, Struyf A, Hubert M, and Hornik K (2015) R package cluster: Cluster Analysis Basics and Extensions. Retrieved from https://cran.r-project.org/web/packages/cluster/cluster.pdf

Marmion W (2015) Evaluating and improving the SAMA (Segmentation Analysis and Market Assessment) recruiting model (Master's thesis). Retrieved from Calhoun http://calhoun.nps.edu/bitstream/handle/10945/45894

Moffit, JD (2016) Analysis of regional effects on market segment production (Unpublished master's thesis) Personal communication May 9, 2016.

Navy Recruiting Command (NRC) (2009) COMNAVCRUITCOMINST 1130.8H – VOLUME I: Recruiting operations. (NRC). Retrieved Feb2 25, 2016

Navy Recruiting Command (NRC) (2016a) Leads data. Unpublished dataset.

Navy Recruiting Command (NRC) (2016b) Navy recruiting facts and statistics. Accessed April 26, 2016, http://www.cnrc.navy.mil/facts-and-stats.htm.

Nielsen (2016) My best segments. Retrieved from Nielsen https://segmentationsolutions.nielsen.com/mybestsegments/

Parker N (2015) Improved Army Reserve unit stationing using market demographics. (Master's thesis) Retrieved from Calhoun http://calhoun.nps.edu/handle/10945/45921

Pinelis Y K, Schmitz E, Miller Z, Rebhan E (2011) An analysis of Navy recruiting goal allocation model. Center For Naval Analysis, Arlington, VA, https://www.cna.org/CNA_files/PDF/D0026005.A2.pdf.

R Development Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing.

Sailer P and Weber M (1998) The IRS population count: An update. Statistics of Income (SOI) working papers. Retrieved from https://www.irs.gov/pub/irs-soi/indpopct.pdf

Smith C and Smith M (2016) LA HomeTownLocator: New Orleans, LA 70176 ZIP code profile. Retrieved from http://louisiana.hometownlocator.com/zip-codes/data,zipcode,70176.cfm

United States Army Recruiting Command (USAREC) (2016) PRIZM NE segments. Provided via J.D. Moffit via personal communication May 10, 2016.

United States Census Bureau (2010) 2010 FIPS codes for counties and county equivalent entities. Accessed March 24, 2016, https://www.census.gov/geo/reference/codes/cou.html

United States Census Bureau (Census) (2016) Population Estimates Terms and Definitions. Accessed February 8, 2016, http://www.census.gov/popest/about/terms.html

United States Government Accountability Office (GAO) (2003) DOD needs to establish objectives and measures to better evaluate advertising's effectiveness. Report, U.S. Government Accountability Office, Washington, DC.

United States Government Accountability Office (GAO) (2010) Military Recruiting: clarified reporting requirements and increased transparency could strengthen oversight over recruiter irregularities. Report, U.S. Government Accountability Office, Washington, DC.

United States Postal Service (USPS) (2015) Management alert – ZIP code review process (Report Number MS-MT-16-001). Retrieved from https://www.uspsoig.gov/sites/default/files/document-library-files/2015/MS-MT-16-001_0.pdf

Williams T (2014) Understanding factors influencing Navy recruiting production. (Master's thesis) Retrieved from Calhoun http://calhoun.nps.edu/handle/10945/48125

Woods & Poole (2013) Long-term county forecasts of employment, population, income, retail sales & households. Retrieved from Woods & Poole http://www.woodsandpoole.com/

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California